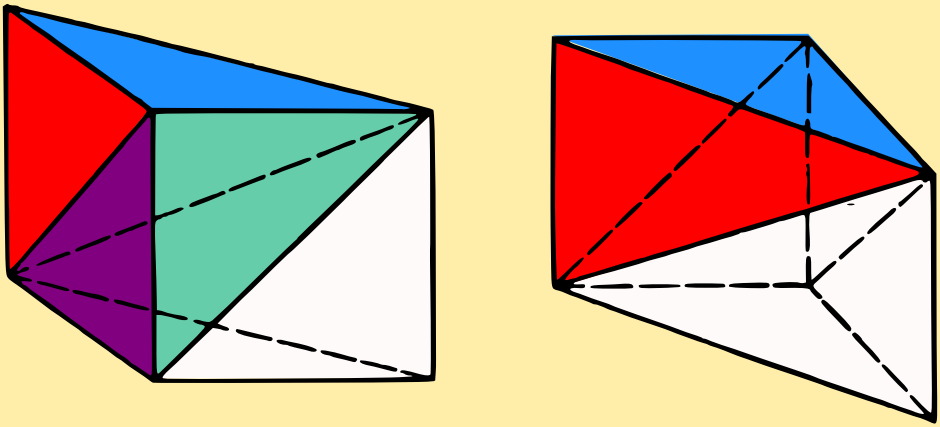
G. Marchouk

METHODES DE CALCUL NUMERIQUE



Éditions Mir Moscou

Г. И. МАРЧУК

МЕТОДЫ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА

G. MARCHOUK

MÉTHODES DE CALCUL NUMÉRIQUE

ÉDITIONS MIR • MOSCOU

Traduit du russe
par I. PÉTROVA

На французском языке

© Главная редакция физико-математической литературы
издательства « Наука » 1977
© Traduction française Editions Mir 1980

PRÉFACE

Le contenu de ce livre développe un cours de Calcul numérique fait à la Faculté de mathématiques de l'Université de Novossibirsk. L'auteur étudie en premier lieu les problèmes complexes de la physique mathématique qui se ramènent en général à des problèmes plus simples, bien étudiés sur le plan théorique et justiciables d'algorithmes efficacement traités sur ordinateurs. Ces problèmes se posent fréquemment aux chercheurs néophytes, et le but principal de ce livre est de leur donner de bonnes recettes en la matière.

Désireux de gagner au Calcul numérique un vaste public, l'auteur a consenti à certaines omissions pour s'appesantir en revanche sur des idées et des tactiques fondamentales. Quant aux points secondaires, quelquefois importants, et à des généralisations possibles telles que les conditions minimales de régularité des fonctions, les contraintes sur les données initiales, etc., ils sont évidents pour un spécialiste et constituent autant d'exercices utiles pour un débutant.

Le Chapitre 9 est issu du rapport fait par l'auteur à Nice, au Congrès international des mathématiciens de 1970. Il fait une sorte de synthèse des méthodes et des questions examinées dans le corps de l'ouvrage, ainsi que d'autres directions de l'Analyse numérique dont l'importance est grande aussi bien du point de vue théorique que pour les applications.

Bien que ce livre emprunte une partie des résultats à la monographie du même nom parue en russe en 1973, il diffère de celle-ci de façon notable. Il expose plusieurs idées et algorithmes nouveaux qui présentent un intérêt certain, méthodique et pratique. On y examine par exemple des algorithmes d'optimisation basés sur les principes variationnels, l'automatisation des calculs moyennant la méthode du domaine auxiliaire, un algorithme itératif de décomposition pour les opérateurs non commutatifs, la méthode de factorisation incomplète. L'interpolation à l'aide de fonctions splines constitue désormais un chapitre à part, ainsi que l'extrapolation de Richardson pour des schémas de haute précision. Les pages traitant des méthodes variationnelles des différences finies sont enrichies de

plusieurs résultats nouveaux tels que la représentation des fonctions continues par des bases discontinues par morceaux et la construction de bases qui tiennent compte des singularités des solutions. Le chapitre consacré à la résolution des problèmes inverses puise plus abondamment dans la théorie des perturbations en rapport avec les problèmes non linéaires de la physique mathématique et dans l'analyse de modèles mathématiques sensibles aux données initiales. La liste des additions ne s'arrête pas là-dessus. A notre avis, l'ouvrage ainsi complété donne une idée plus juste des méthodes de Calcul numérique pour les problèmes complexes de la science et de la technique.

L'auteur adresse ses profonds remerciements à Y. Kouznétsov dont la contribution à ce livre est impossible à surestimer. Il est fort obligé à V. Smélov, V. Iline, V. Chaïdourov, V. Agochkov, A. Matsokine, G. Rivine, V. Tsétsokho, V. Vassilenko, ainsi qu'à V. Drobychévitch, V. Dymnikov et V. Pénenko qui lui ont été d'une grande aide au cours de la rédaction de ce livre.

N. Bakhvalov, M. Fagué et V. Lébédev ont lu le manuscrit et fait plusieurs remarques de valeur. MM. Lions et Glowinski ont eu l'obligeance de mettre à la disposition de l'auteur des textes traitant des inéquations variationnelles. Qu'ils veuillent bien trouver ici l'expression de sa respectueuse gratitude.

G. Marchouk

INTRODUCTION

L'avènement des calculateurs électroniques modernes a accru considérablement les possibilités des chercheurs dans la simulation mathématique des problèmes complexes de la science et de la technique. On comprend donc qu'à l'heure actuelle les méthodes de recherche quantitatives débordent sur toutes les activités humaines et que les modèles mathématiques deviennent un instrument de connaissance puissant.

L'ambition des modèles mathématiques ne se borne pas, loin de là, à faire connaître les lois. Leur rôle s'accroît avec la tendance naturelle à optimiser les équipements et la technologie de la planification des expériences. Désireux d'approfondir ses connaissances et de se faire une idée juste des phénomènes étudiés, l'homme conçoit la nécessité de modèles mathématiques toujours plus fins qui demandent à leur tour un appareil mathématique ingénieux et passe-partout. Les modèles sont traités sur ordinateur par les méthodes du Calcul numérique qui évolue sans cesse avec le hardware.

Toute réduction des problèmes de la physique mathématique et de la technique conduit en fin de compte à des équations algébriques de diverses formes, si bien que l'objet du Calcul numérique tourne en général autour des procédés de réduction et des méthodes de résolution des systèmes algébriques ainsi obtenus.

La formation du système algébrique associé à un problème à arguments variant de façon continue s'appuie d'ordinaire sur l'information à priori relative au problème proposé. Cela peut être entre autres l'appartenance de la solution à une classe de fonctions possédant une certaine régularité, les propriétés des opérateurs et des données initiales. L'information à priori dicte souvent le choix de la méthode numérique destinée à résoudre les équations algébriques, et les conditions à priori imposées au problème initial doivent en général être compatibles avec les propriétés de l'analogue algébrique. Cela est surtout vrai des opérateurs dont la discrétisation doit conserver au possible les propriétés.

Il paraît que ce principe gouverne la résolution de nombreux problèmes. On note d'ailleurs que la « filiation » opératorielle men-

tionnée permet d'utiliser des méthodes bien élaborées de l'Analyse fonctionnelle et d'étudier donc l'efficacité des algorithmes numériques par des procédés universels simples.

Nous donnons ci-dessous un bref résumé du livre en mettant en relief les résultats principaux et les idées originales.

Le *Chapitre premier* est consacré aux éléments de la théorie des schémas aux différences. On y trouve des notions classiques (approximation, stabilité numérique, convergence des solutions) et des résultats importants relatifs aux propriétés générales des problèmes et de leurs adjoints, dont on aura besoin dans d'autres chapitres. Nous attirons l'attention du lecteur sur le point 1.1.2 qui décrit des algorithmes actuellement utilisés pour calculer les bornes du spectre non négatif des matrices. Il est connu que la borne supérieure du spectre est définie par des procédés itératifs bien élaborés et que le traitement numérique correspondant ne présente d'habitude aucune difficulté. Quant au calcul de la plus petite valeur propre (la borne inférieure du spectre), il s'agit là d'un problème ardu.

Théoriquement, la tactique la plus simple consiste à évaluer la plus grande valeur propre de l'opérateur inverse et conduit à des algorithmes médiocres. L'auteur expose un procédé avec translation qui fournit assez facilement la borne inférieure cherchée. Si nous insistons sur cette question, c'est que de nombreux algorithmes de calcul, particulièrement des procédures liées à l'optimisation des processus itératifs, s'inspirent de l'information à priori sur les bornes du spectre.

Le *Chapitre 2* étudie les procédés de construction de schémas aux différences. Il est centré autour de la méthode des relations intégrales et des procédés variationnels. Chacune de ces tactiques a ses mérites et ses inconvénients. Elles ne sont pas autonomes : dans un contexte déterminé, on aboutit dans les deux cas aux mêmes schémas aux différences qui approchent les problèmes différentiels proposés.

Nombreux sont néanmoins les cas où l'on préfère l'approche variationnelle car elle conserve les propriétés de définitivité des opérateurs. Chose importante, cela a lieu automatiquement pour une classe étendue de problèmes.

En ce qui concerne les techniques variationnelles, nous nous limitons aux méthodes de Ritz, de Galerkin et des moindres carrés. Ces méthodes n'épuisent certes pas l'arsenal variationnel, mais elles permettent de se familiariser avec les idées générales qui président à la construction de schémas aux différences et se transposent facilement à d'autres cas.

Quelques mots au sujet de la méthode des éléments finis. Il s'agit là d'un outil commode qui s'apparente sur le plan méthodologie à la méthode de Fourier à la différence qu'on opère non pas avec les bases de fonctions usuelles (par exemple les bases de fonctions trigonométriques, de polynômes de Legendre, de polynômes d'Her-

mite), mais avec des polynômes partout nuls, sauf dans une région assez petite qui est le domaine de variation des arguments, d'où le nom « éléments finis ».

Les emprunts au Calcul variationnel ont leur raison d'être. En effet, on sait qu'une fonctionnelle variationnelle qui traduit fidèlement certaines lois de la mécanique, de la physique mathématique, de la dynamique, atteint son extrémum sur la solution du problème proposé. Si l'on se donne cette fonctionnelle et si l'on définit la classe de fonctions susceptibles de la minimiser, le problème se ramène à la recherche algorithmique de la fonction minimisante.

Si l'on restreint la classe de fonctions admissibles par le jeu de contraintes supplémentaires, cette fonction n'est pas forcément la solution du problème primitif, elle n'en est qu'une approximation.

Le rôle des fonctionnelles variationnelles en physique mathématique augmentera sans cesse avec la puissance du matériel de calcul. On mettra au point des procédés d'exploration dirigée des fonctions d'essai appartenant à de vastes classes, qui permettront la recherche efficace des solutions extrémales. On débouche ici sur le problème d'organiser de façon optimale le calcul pour obtenir une solution avec une précision désirée, i.e. sur la théorie de l'optimisation.

Certains problèmes de la science et de la technique ne se posent pas de façon classique. Les problèmes avec contraintes sont du nombre encore que les cas simples soient classiques. Les contraintes sont par exemple les conditions aux limites pour les problèmes différentiels.

Dans les cas plus complexes, on fait appel à un appareil mathématique plus fin. Si l'on demande par exemple la flexion d'une membrane soumise à diverses forces et si la position de la pièce est majorée et minorée par des fonctions de coordonnées données, l'approche classique s'avère inopérante. Mais si l'on associe à ce problème une fonctionnelle variationnelle dont on cherche le minimum dans la classe de fonctions vérifiant la contrainte donnée, alors la fonction minimisante fournit la solution cherchée.

De nombreux résultats dans ce domaine sont dus aux mathématiciens français qui ont introduit les inéquations variationnelles spéciales aux problèmes avec contraintes. Ces questions font l'objet du n° 2.8.

Dans le *Chapitre 3* on considère l'interpolation des fonctions discrètes. Le problème d'interpolation intervient toutes les fois qu'on étend une fonction définie sur un réseau par des fonctions continues au domaine tout entier. On y rapporte le problème de prolonger une solution approchée à tout le domaine connaissant ses valeurs aux nœuds du réseau et cet autre qui consiste à traiter les données expérimentales associées à un ensemble ponctuel discrétisé.

L'interpolation devient la clef de voûte de l'automatisation des calculs de bureaux d'étude qui suggèrent par leur nature même les

procédés de représentation graphique de l'information. Le concept d'interpolation n'est pas nouveau et les méthodes classiques correspondantes ont leur place dans la littérature mathématique. Aussi nous nous occuperons de cet être mathématique de création récente qu'est la fonction spline.

L'interpolation à l'aide de fonctions splines est le meilleur moyen de prolonger de façon régulière les fonctions discrètes en des fonctions de classes données. Le caractère optimal des fonctions splines tient à leur propriété extrémale spéciale. Selon nous, le lecteur doit être au fait de l'approximation spline dont l'utilisation se répand dans tous les domaines scientifiques et techniques.

Le *Chapitre 4* étudie surtout les méthodes itératives pour les équations algébriques linéaires. On y trouve aussi bien des procédés généraux que des techniques particulières aux problèmes approchés par les différences finies ou par les méthodes variationnelles des différences. Malgré la profusion de travaux décrivant les méthodes itératives et les algorithmes efficaces, nous avons jugé utile d'insister, sans oublier toutefois les procédés classiques, sur les techniques itératives qu'on optimise moyennant les fonctionnelles quadratiques. Telle est en général notre façon de traiter le problème d'optimisation qu'il s'agisse de construire les algorithmes de calcul ou de les exécuter sur ordinateur.

Pour ce qui est des cas spéciaux aux matrices de forme particulière qui interviennent dans la résolution numérique des problèmes de la physique mathématique, nous recommandons de décomposer les matrices en des matrices plus simples au sein du processus itératif. La méthode de décomposition continue naturellement la méthode des directions alternées qui a joué un rôle exceptionnel en physique mathématique. Elle admet diverses variantes et généralisations dont certaines utilisent les principes variationnels.

On note l'intérêt des méthodes directes de résolution des équations aux différences finies telles que la transformation de Fourier rapide et la réduction cyclique. Ces méthodes s'implantent de plus en plus résolument en calcul numérique.

Les méthodes de résolution des problèmes non stationnaires font l'objet du *Chapitre 5*. Elles procèdent essentiellement de la décomposition d'opérateurs. On analyse la stabilisation et la prédiction-correction, techniques qui ont déjà fait leurs preuves, et on décrit en détail la méthode de splitting-up qui est, selon nous, un procédé des plus efficaces (pour l'idée relative à cette méthode voir point 5.3.3 et numéro 5.4).

La méthode de splitting-up ramène à chaque pas temporel un grand problème de la physique mathématique à une suite de problèmes élémentaires. Elle donne lieu à un algorithme efficacement applicable sur ordinateur. Cet algorithme est stable et garantit une solution approchée d'ordre 2 par rapport à toutes les variables. La méthode

en question s'applique à une nombreuse classe de problèmes non stationnaires.

Le *Chapitre 6* examine les méthodes d'amélioration de la précision qui remontent à Richardson et à Runge. Il est connu qu'on affine les solutions de diverses manières. On cherche d'ordinaire une meilleure approximation des équations différentielles ou intégrales à l'aide de schémas de haute précision. Richardson a proposé d'utiliser une approximation aux différences peu précise et de l'associer à des réseaux différents. Si l'équation discrétisée de départ correspond à un réseau de pas h, l'équation suivante correspond au pas $h/2$, et ainsi de suite. On obtient une suite d'équations relatives à des réseaux successifs. Il se trouve que si l'opérateur et les données initiales satisfont à certaines conditions, la combinaison linéaire de résultats obtenus avec plusieurs pas différents fournit une solution plus exacte que les facteurs.

L'extrapolation de Richardson a été primitivement une méthode pour les équations différentielles ordinaires. Aujourd'hui, on l'applique également aux problèmes aux limites relatifs aux équations elliptiques et paraboliques. Ces cas présentent naturellement certaines particularités qui sont reflétées par les schémas de mise en œuvre numérique. Chose importante, le procédé de Richardson joue pour les problèmes avec paramètre arbitraire et pour les problèmes conditionnellement bien posés abordés par régularisation. Il procède alors de la résolution de problèmes avec paramètres différents qui tendent vers une valeur limite. Ainsi, la méthode considérée apporte en Analyse numérique des idées nouvelles qui se montrent très fructueuses dans l'optimisation d'algorithmes de calcul.

S'agissant des problèmes traités par les méthodes variationnelles des différences, l'extrapolation de Richardson s'acquitte d'une mission d'importance. En effet, on se trouve en général devant l'alternative : ou bien résoudre l'équation aux différences associée à un pas très petit moyennant des approximations assez rudes, ou bien l'aborder avec des pas plus grands à l'aide des schémas de haute précision.

Le premier procédé est simple, mais il exige un volume de calcul important, et le second est logiquement beaucoup plus difficile tout en demandant moins d'opérations arithmétiques. Ainsi, quand on veut un résultat plus précis, ils sont inopérants l'un et l'autre. On a donc proposé d'utiliser des schémas variationnels aux différences simples qui n'approchent qu'à l'ordre 1 ou 2 et de les associer à une suite de réseaux. Dans de nombreux cas, la combinaison linéaire de solutions ainsi obtenues fournit, on l'a déjà dit, une solution avec une précision voulue.

Le *Chapitre 7* examine la position et la résolution numérique des problèmes inverses. Dans les méthodes de simulation mathématique de grands problèmes de la science et de la technique, on doit

souvent reconstituer la solution par des fonctionnelles ou reconstituer un opérateur de forme déterminée. Cette classe de problèmes inverses s'avère la plus laborieuse sur le plan numérique puisqu'elle est en général liée à la résolution de problèmes mal posés au sens d'Hadamard. Ces derniers problèmes ont ouvert un nouveau chapitre des mathématiques. Les résultats les plus importants dans ce domaine sont dus à l'école mathématique soviétique. Après que Tikhonov a introduit la régularisation des problèmes mal posés, on a eu vite fait de les justifier et de mettre sur pied l'outillage algorithmique correspondant. Ce chapitre est centré sur la position des problèmes inverses qui consistent à reconstituer les opérateurs différentiels et les données initiales. Car si l'on fixe la forme d'un opérateur différentiel, ses coefficients sont néanmoins des inconnues à déterminer. La théorie des problèmes inverses emprunte beaucoup aux équations adjointes. L'appareil mathématique élaboré par l'auteur s'avère efficace lorsqu'on se propose d'évaluer les petites perturbations des fonctionnelles de la solution en fonction des variations d'entrées. Il y a lieu de dire que la théorie des perturbations élaborée dans ce chapitre s'applique aussi bien aux problèmes linéaires qu'à ceux qui ne le sont pas.

Le *Chapitre 8* examine certaines méthodes décrites antérieurement en rapport avec des problèmes concrets de la physique mathématique.

Le *Chapitre 9* fait le tour des méthodes d'Analyse numérique. Ce livre n'épuise certes pas toute la diversité d'algorithmes de calcul connus et il en passe beaucoup sous silence parce qu'ils sont décrits à fond dans des manuels et des ouvrages spécialisés. C'est le cas par exemple des éléments classiques de l'Analyse numérique tels que les formules de cubature, les méthodes pour les équations différentielles ordinaires, l'interpolation usuelle. Quant aux procédés de pointe, le chapitre 9 donne une idée de quelques-uns d'entre eux (Monte-Carlo, grosses particules, relations intégrales, méthodes universelles d'algèbre linéaire).

Il initie de plus à certains problèmes qui se présentent en Analyse numérique et examine les voies de recherche correspondantes. Selon nous, cela aidera le lecteur à s'orienter dans cette discipline et à en déterminer les secteurs en plein essor.

Ce livre est un manuel, si bien que nous avons réduit au possible (sauf au Chapitre 9) les références à la Bibliographie *in fine*.

La *Bibliographie* est conçue de façon que le lecteur entre vite dans le vif de la question qui l'intéresse.

L'auteur tient à dire pour conclure que le but de ce livre est de former le lecteur à la résolution numérique des problèmes complexes de la science et de la technique.

CHAPITRE PREMIER

THÉORIE DES SCHÉMAS AUX DIFFÉRENCES. GÉNÉRALITÉS

Ce chapitre donne des notions fondamentales en théorie des schémas aux différences qui seront l'assise des chapitres suivants. Notre but principal étant d'initier le lecteur à certains principes modernes de la construction d'algorithmes de calcul pour les problèmes de la physique mathématique, nous nous bornerons, en matière de la théorie, à des cas simples.

1.1. Opérateurs et leurs adjoints

Considérons dans l'espace euclidien E_n de dimension n un domaine D. Les mathématiques appliquées n'emploient en général pas de domaines complexes au point d'être dépourvus de mesure (de mesure de surface en dimension deux, de mesure de volume en dimension trois, etc.). Comme la théorie de la mesure de Lebesgue est à la base des définitions ultérieures, le lecteur est supposé connaître cette théorie et la théorie de l'intégrale de Lebesgue (Smirnov [II], Sobolev [I; 6], Wladimirov [II], Natanson [I; 4]) *).

Appelons $L_2(D)$ l'espace hilbertien de toutes les fonctions mesurables réelles $f(x)$ de carré sommable, i.e. telles que

$$\int_D f^2(x)\,dx < \infty,$$

muni du produit scalaire

$$(f, g) = \int_D f(x)\, g(x)\, dx. \tag{1.1}$$

Les fonctions f de $L_2(D)$ sont normées de façon usuelle par l'égalité

$$\| f \|^2 = (f, f)^{1/2}. \tag{1.2}$$

On extrait de $L_2(D)$ un sous-espace (une variété linéaire) $\Phi \subset \subset L_2(D)$ tel que chaque élément $\varphi \in \Phi$ vérifie certaines conditions

*) Le chiffre romain désigne la section de la Bibliographie *in fine* et le chiffre arabe indique, si besoin est, le numéro à l'intérieur de la section.

supplémentaires, par exemple une condition de régularité, la propriété de satisfaire sur la frontière de D à des relations limites, et ainsi de suite. Ces conditions doivent être suffisantes pour qu'un opérateur A transforme $\varphi \in \Phi$ en $A\varphi \in L_2(D)$.

Un opérateur linéaire A défini sur la variété linéaire Φ est dit *semi-défini positif* si

$$(A\varphi, \varphi) \geqslant 0 \tag{1.3}$$

pour tout $\varphi \in \Phi$ et s'il existe $\varphi \neq 0$ pour lesquels $(A\varphi, \varphi) = 0$. La notation employée est $A \geqslant 0$. Si

$$(A\varphi, \varphi) > 0 \tag{1.4}$$

pour $\varphi \neq 0$, l'opérateur A s'appelle *positif* et se note $A > 0$. S'agissant de la relation plus forte

$$(A\varphi, \varphi) \geqslant \gamma (\varphi, \varphi), \tag{1.5}$$

avec $\gamma > 0$ une constante positive, la même pour tous les $\varphi \in \Phi$, l'opérateur A est dit *défini positif*.

Si A est une matrice carrée d'ordre fini, sa positivité entraîne la définie positivité (Faddeev, Faddeeva [VIII]).

Nous dirons que Φ constitue le *domaine de l'opérateur* A et nous le désignerons par $\Phi(A)$.

On introduit par l'identité de Lagrange

$$(Ag, h) = (g, A^*h), \tag{1.6}$$

avec $g \in \Phi(A)$, $h \in \Phi(A^*)$, l'*adjoint* A^* de A.

Les sous-espaces $\Phi(A)$ et $\Phi(A^*)$ de l'espace de Hilbert $L_2(D)$ ne coïncident en général pas bien que leurs éléments soient de règle définis dans un même domaine.

Si $A = A^*$ et $\Phi(A) \equiv \Phi(A^*)$, l'opérateur A est dit *auto-adjoint*.

Les propriétés des opérateurs adjoints donnent lieu à un résultat important: si $\Phi(A) \equiv \Phi(A^*)$, alors $A > 0$ entraîne $A^* > 0$.

Les algorithmes utilisent souvent le développement des fonctions en séries de Fourier suivant les fonctions propres d'opérateurs et de leurs adjoints.

Considérons deux problèmes spectraux pour $A \geqslant 0$:

$$Au = \lambda(A)\, u, \quad A^*u^* = \lambda(A^*)\, u^*. \tag{1.7}$$

On suppose que chaque équation homogène (1.7) forme un système complet de fonctions propres $\{u_n\}$ et $\{u_n^*\}$ normées par

$$(u_n, u_m^*) = \begin{cases} 1, & n = m, \\ 0, & n \neq m, \end{cases} \tag{1.8}$$

et que les valeurs propres associées $\lambda_n(A)$ et $\lambda_n(A^*)$ sont réelles. Il est classique que $\lambda_n(A) = \lambda_n(A^*)$. Admettons que les valeurs

propres des problèmes proposés appartiennent à l'intervalle $[\alpha, \beta]$:

$$\alpha \leqslant \lambda_n (A) \leqslant \beta.$$

Ce système complet de fonctions propres constitue une *base biorthonormée*. Sous l'hypothèse de complétude, toute f de Φ et toute f^* de Φ^* s'écrivent sous forme de séries de Fourier *)

$$f = \sum_n f_n u_n, \qquad f^* = \sum_n f_n^* u_n^*, \tag{1.9}$$

où

$$f_n = (f, u_n^*), \qquad f_n^* = (f^*, u_n). \tag{1.10}$$

En Analyse numérique, on recourt souvent à l'évaluation des normes d'opérateurs. La *norme de l'opérateur* A se définit par

$$\| A \|^2 = \sup_{\substack{\varphi \in \Phi \\ \varphi \neq 0}} \frac{(A\varphi, A\varphi)}{(\varphi, \varphi)} \tag{1.11}$$

(dans la suite on omettra la contrainte $\varphi \neq 0$). Etant donné

$$(A\varphi, A\varphi) = (\varphi, A^*A\varphi),$$

(1.11) se récrit

$$\| A \|^2 = \sup_{\varphi \in \Phi} \frac{(\varphi, A^*A\varphi)}{(\varphi, \varphi)}. \tag{1.12}$$

L'opérateur A^*A est semi-défini positif et symétrique. Considérons le problème spectral

$$A^*A\Omega = \lambda (A^*A)\, \Omega \tag{1.13}$$

qui détermine le système de fonctions propres orthonormées $\{\Omega_n\}$ et de valeurs propres $\lambda_n (A^*A) \geqslant 0$. Le système $\{\Omega_n\}$ est supposé complet. La fonction φ se met sous forme de série de Fourier

$$\varphi = \sum_n \varphi_n \Omega_n, \tag{1.14}$$

où

$$\varphi_n = (\varphi, \Omega_n). \tag{1.15}$$

Portons (1.14) dans (1.12), il vient, compte tenu de la condition d'orthonormalité des Ω_n,

$$\| A \|^2 = \sup_{\{\varphi_n\} \in Q} \frac{\sum_n \lambda_n (A^*A)\, \varphi_n^2}{\sum_n \varphi_n^2}, \tag{1.16}$$

*) Par souci de simplicité, on écrira Φ et Φ^* au lieu de $\Phi(A)$ et $\Phi(A^*)$.

Q étant l'espace des coefficients de Fourier. On vérifie aisément que

$$\frac{1}{\| A^{-1} \|^2} = \lambda_{\min}(A^*A) = \alpha(A^*A),$$
$$\| A \|^2 = \lambda_{\max}(A^*A) = \beta(A^*A), \tag{1.17}$$

où $\lambda_{\min}(A^*A)$ et $\lambda_{\max}(A^*A)$ désignent respectivement la plus petite et la plus grande valeur propre de l'ensemble $\{\lambda_n(A^*A)\}$ du problème spectral (1.13). La quantité $\beta(A^*A) = \lambda_{\max}(A^*A)$ s'appelle d'ordinaire le *rayon spectral* de l'opérateur A^*A, et le rayon spectral de A est $\beta(A) = \sup\{ |\lambda(A)| \}$.

Soit A un opérateur auto-adjoint. Considérons le problème spectral

$$Au = \lambda u. \tag{1.18}$$

On a

$$\| A \| = \beta(A). \tag{1.19}$$

La propriété de A d'être auto-adjoint entraîne de suite

$$\beta(A^2) = [\beta(A)]^2. \tag{1.20}$$

Soit donné, dans $L_2(D)$, un opérateur positif C. On a

$$(C\varphi, \varphi) > 0 \tag{1.21}$$

pour tous les $\varphi \neq 0$ de Φ *partout dense* dans $L_2(D)$, i.e. si $f \in L_2(D)$ et $g \in \Phi$, il existe pour tout f un élément g tel qu'on ait $\| f - g \| \leqslant \varepsilon$, ε étant une constante positive arbitrairement petite. Admettons que le domaine Φ^* de C^* auto-adjoint ne diffère pas de Φ, si bien qu'on trouve $C^*\varphi$ pour tout $\varphi \in \Phi$. Dans ce cas $(C^*\varphi, \varphi) = (\varphi, C\varphi) = (\bar{C}\varphi, \varphi)$, où $\bar{C} = \frac{1}{2}(C + C^*)$. Comme C est positif symétrique, on munit Φ du nouveau produit scalaire $(f, g)_C = (Cf, g)$ et de la norme

$$\| \varphi \|_C^2 = (C\varphi, \varphi) = (\bar{C}\varphi, \varphi),$$

dite *norme énergie* ou C-norme. On obtient l'estimation importante

$$\| \varphi \|_C^2 = \| \varphi \|_{\bar{C}}^2 \leqslant \| \bar{C} \| \, \| \varphi \|^2 = \beta(\bar{C}) \, \| \varphi \|^2, \tag{1.22}$$

avec $\beta(\bar{C})$ la plus grande valeur propre de l'opérateur $\bar{C}$.

Il nous reste à noter l'intérêt des fonctions de l'espace de Sobolev $W_2^l(D)$. Il s'agit d'un espace de Hilbert ayant pour éléments les fonctions $\in L_2(D)$ qui possèdent dans D des dérivées généralisées premières de carré sommable. On définit dans cet espace le produit scalaire par la formule (voir Sobolev [I; 6], Wladimirov [II])

$$(u, v)_{W_2^l} = \sum_{k=0}^{l} \sum_{(k)} \int_D \frac{\partial^k u}{\partial x^k} \frac{\partial^k v}{\partial x^k} \, dD, \tag{1.23}$$

les dérivées étant désignées par

$$\frac{\partial^k \varphi}{\partial x^k} = \frac{\partial^{\alpha_1+\ldots+\alpha_n}}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} \varphi, \qquad \alpha_1 + \ldots + \alpha_n = k.$$

La norme sur $W_2^l(D)$ est définie par

$$\| \varphi \|_{W_2^l} = (\varphi, \varphi)_{W_2^l}^{1/2}. \tag{1.24}$$

Si φ sont dans l'espace de Sobolev $W_2^l(D)$ et remplissent la condition

$$\varphi = 0 \text{ sur la frontière } \partial D \text{ de } D,$$

on emploie la notation $\mathring{W}_2^l(D)$.

1.1.1. Estimation des normes de matrices

Soit $A \geqslant 0$ une matrice semi-définie positive agissant sur les vecteurs de l'espace euclidien. On a

$$\| (E + \sigma A)^{-1} \| \leqslant 1 \tag{1.25}$$

(E est la matrice unité) quel que soit le paramètre $\sigma \geqslant 0$. La démonstration utilise la formule

$$\| (E + \sigma A)^{-1} \|^2 = \max_{\varphi} \frac{((E+\sigma A)^{-1} \varphi, (E+\sigma A)^{-1} \varphi)}{(\varphi, \varphi)}. \tag{1.26}$$

Posons

$$\psi = (E + \sigma A)^{-1} \varphi,$$

auquel cas

$$\| (E + \sigma A)^{-1} \|^2 = \max_{\psi} \frac{(\psi, \psi)}{((E+\sigma A)\psi, (E+\sigma A)\psi)} =$$

$$= \frac{1}{\min\limits_{\psi} \left[1 + 2\sigma \dfrac{(A\psi, \psi)}{(\psi, \psi)} + \sigma^2 \dfrac{(A\psi, A\psi)}{(\psi, \psi)} \right]}.$$

Puisque $A \geqslant 0$, il en résulte l'estimation (1.25) pour les vecteurs φ et ψ. Lorsque $A > 0$, on a évidemment pour $\sigma > 0$

$$\| (E + \sigma A)^{-1} \| < 1. \tag{1.27}$$

LEMME (KELLOGG [XV]). *Si la matrice* $A \geqslant 0$ *et* $\sigma \geqslant 0$, *alors*

$$\| (E - \sigma A)(E + \sigma A)^{-1} \| \leqslant 1. \tag{1.28}$$

Introduisons la notation

$$T = (E - \sigma A)(E + \sigma A)^{-1}$$

et considérons l'expression de $\|T\|^2$

$$\|T\|^2 = \max_{\varphi} \frac{((E-\sigma A)(E+\sigma A)^{-1}\varphi, (E-\sigma A)(E+\sigma A)^{-1}\varphi)}{(\varphi, \varphi)} =$$

$$= \max_{\psi} \frac{((E-\sigma A)\psi, (E-\sigma A)\psi)}{((E+\sigma A)\psi, (E+\sigma A)\psi)} =$$

$$= \max_{\psi} \frac{(\psi, \psi) - 2\sigma(A\psi, \psi) + \sigma^2(A\psi, A\psi)}{(\psi, \psi) + 2\sigma(A\psi, \psi) + \sigma^2(A\psi, A\psi)} \leqslant 1.$$

Le fait que A est semi-définie positive est en l'occurrence essentielle.

Pour A positive, on a avec $\sigma > 0$:

$$\|(E - \sigma A)(E + \sigma A)^{-1}\| < 1. \tag{1.29}$$

1.1.2. Recherche des bornes du spectre d'une matrice positive

Soit le problème de trouver les valeurs propres extrêmes d'une matrice $A > 0$ à spectre positif. On utilise la méthode de Ljusternik [IV; 16].

Soit le problème spectral

$$Au = \lambda u. \tag{1.30}$$

On suppose qu'il détermine le système complet de vecteurs propres $u_k \in \Phi$ et le système de valeurs propres positives $\lambda_k(A)$. (Le problème du spectre est étudié à fond par Marek [VIII; 11, 30] et Faddeev, Faddeeva [VIII].) Considérons le processus itératif

$$\varphi^{(n+1)} = \frac{1}{c_n} A\varphi^{(n)},$$

$$\varphi^{(0)} = g,$$

avec g un vecteur non nul quelconque et c_n un facteur de normalisation choisi de préférence sous la forme

$$c_n = \|\varphi^{(n)}\|.$$

On a

$$\varphi^{(n+1)} = A\frac{\varphi^{(n)}}{\|\varphi^{(n)}\|}. \tag{1.31}$$

Soit $0 < \alpha(A) = \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_m = \beta(A)$. Evidemment

$$\beta(A) = \lim_{n\to\infty} \|\varphi^{(n)}\|. \tag{1.32}$$

La complétude supposée du système de vecteurs u_n entraîne en effet le développement

$$\varphi^{(0)} = \sum_k g_k u_k,$$

où $g_k = (g, u_k^*)$, $\{u_k^*\}$ étant un système de vecteurs propres de la matrice A^*. Vu la relation de récurrence

$$\varphi^{(n)} = A\frac{\varphi^{(n-1)}}{\| \varphi^{(n-1)} \|} = \quad . \quad = \frac{A^n g}{\| A^{n-1} g \|},$$

on a

$$\lim_{n\to\infty} \| \varphi^{(n)} \| = \lim_{n\to\infty} \frac{\| A^n g \|}{\| A^{n-1} g \|}.$$

Puisque

$$A^n g = \sum_k [\lambda_k (A)]^n g_k u_k,$$

on obtient pour n suffisamment grands

$$A^n g = \beta^n (A)\, g_m \left\{ u_m + O\left[\left(\frac{\lambda_{m-1}}{\lambda_m} \right)^n \right] \right\},$$

avec $\beta (A) = \lambda_m$ la plus grande valeur propre de A.

Si l'approximation initiale g est par hasard une combinaison linéaire de vecteurs propres associés aux valeurs propres autres que $\beta (A)$, on trouve toujours $\beta (A)$ par approximations successives car le développement de $\varphi^{(n)}$ renferme toutes les composantes des vecteurs de base (effet des erreurs d'arrondi).

La dernière relation implique

$$\frac{\| A^n g \|}{\| A^{n-1} g \|} = \beta (A) + O\left[\left(\frac{\lambda_{m-1}}{\lambda_m} \right)^n \right],$$

donc

$$\beta (A) = \lim_{n\to\infty} \frac{\| A^n g \|}{\| A^{n-1} g \|}. \tag{1.33}$$

Cherchons la plus petite valeur propre de A. Considérons la matrice *)

$$B = \beta (A)\, E - A \tag{1.34}$$

et le problème spectral

$$Bu = \lambda (B)\, u. \tag{1.35}$$

Evidemment $B \geqslant 0$. Il découle de (1.34) que A et B possèdent une base commune $\{u_k\}$. Imitons le cas précédent et considérons le processus itératif

$$\psi^{(n+1)} = B\frac{\psi^{(n)}}{\| \psi^{(n)} \|}, \tag{1.36}$$

il vient

$$\beta (B) = \lim_{n\to\infty} \| \psi^{(n)} \|. \tag{1.37}$$

*) Au lieu de $\beta (A)$ on peut prendre tout nombre supérieur.

On note que (1.34) et le fait, pour A et B, de posséder une base commune entraînent

$$\beta(B) = \beta(A) - \alpha(A).$$

D'où

$$\alpha(A) = \beta(A) - \beta(B). \tag{1.38}$$

Ainsi, on cherche la plus petite valeur propre des matrices de ce type sous forme de différence de plus grandes valeurs propres de A et $B = \beta(A)E - A$. Comme $\beta(A)$ et $\beta(B)$ s'obtiennent itérativement de façon indiquée, le problème s'avère en principe résolu.

Il y a à noter que pour A mal conditionnées la plus petite valeur propre $\alpha(A)$ est définie comme différence de $\beta(A)$ et $\beta(B)$ importants, si bien que dans cet algorithme les erreurs sont possibles non seulement sur la grandeur de $\alpha(A)$, mais aussi sur son signe. On l'évite en modifiant légèrement l'algorithme. Soit le processus itératif

$$\psi^{(n+1)} = B\psi^{(n)}, \qquad \psi^{(0)} = h.$$

Le système de vecteurs propres u_k de A rangés dans l'ordre de croissance des valeurs propres devient un système ordonné de vecteurs propres v_k de B de façon que $v_k = u_{m-k+1}$, $k = 1, 2, \ldots, m$. Soit

$$A\psi^{(n)} = \sum_{k=1}^{m} \lambda_k^n(B)\, h_k A v_k, \qquad \psi^{(n)} = \sum_{k=1}^{m} \lambda_k^n(B)\, h_k v_k,$$

où $h_k = (h, v_k^*)$. Lorsque $n \gg 1$, on a

$$A\psi^{(n)} \approx \beta^n(B)\, h_m A v_m, \qquad \psi^{(n)} \approx \beta^n(B)\, h_m v_m.$$

Comme $Av_m = Au_1 = \alpha(A)\, u_1 = \alpha(A)\, v_m$, on aboutit à l'algorithme

$$\psi^{(n+1)} = B\frac{\psi^{(n)}}{\|\psi^{(n)}\|}, \tag{1.39}$$

$$\alpha(A) = \lim_{n\to\infty} \frac{\|A\psi^{(n)}\|}{\|\psi^{(n)}\|}. \tag{1.40}$$

Avec la dernière formule, on n'a plus affaire avec la différence de grands nombres, et l'algorithme de recherche des bornes du spectre de l'opérateur A est en général efficacement réalisable sur ordinateur.

On note cependant la convergence lente des processus itératifs (1.31), (1.36) et (1.39). On utilise divers procédés d'accélération dont les plus courants sont la technique de Tchebycheff et les méthodes avec translation (Faddeev, Faddeeva [VIII], Gavourine [IX; 4], Wilkinson [VIII]).

S'agissant des matrices symétriques, le calcul de $\alpha(A)$ et $\beta(A)$ se fait de préférence avec la norme énergie.

Quand on optimise ou effectue des appréciations théoriques des algorithmes, on doit souvent connaître la norme de A et celle de son inverse A^{-1}. On a les relations suivantes vraies pour tout opérateur inversible

$$\| A \|^2 = \sup_{\varphi \in \Phi} \frac{(A\varphi,\ A\varphi)}{(\varphi,\ \varphi)} = \sup_{\varphi \in \Phi} \frac{(A^*A\varphi,\ \varphi)}{(\varphi,\ \varphi)} = \beta\,(A^*A)$$

et

$$\| A^{-1} \|^2 = \sup_{\varphi \in \Phi} \frac{(A^{-1}\varphi,\ A^{-1}\varphi)}{(\varphi,\ \varphi)} = \sup_{\varphi \in \Phi} \frac{((AA^*)^{-1}\,\varphi,\ \varphi)}{(\varphi,\ \varphi)} = [\alpha\,(A^*A)]^{-1}.$$

D'où

$$\| A \| = \sqrt{\beta\,(A^*A)}\,, \tag{1.41}$$

$$\| A^{-1} \| = \left(\sqrt{\alpha\,(A^*A)}\right)^{-1}. \tag{1.42}$$

$\alpha\,(A^*A)$ et $\beta\,(A^*A)$ s'obtiennent par le procédé décrit des approximations successives.

L'algorithme de recherche des bornes du spectre d'une matrice positive permet d'optimiser les processus itératifs par des méthodes très évoluées (voir Chapitre 4). Ces processus deviennent constructifs et s'appliquent heureusement en physique mathématique.

Les bornes du spectre sont bien évaluées (à part la méthode de Ljusternik) par le théorème de Gershgorin (voir Faddeev, Faddeeva [VIII]) : une matrice arbitraire A d'ordre m et d'éléments a_{kl} a toutes ses valeurs propres $\lambda\,(A)$ dans la réunion de cercles

$$| z - a_{kk} | \leqslant R_k, \quad k = 1, 2, \ldots, m,$$

où

$$R_k = \sum_{\substack{l=1 \\ l \neq k}}^{n} | a_{kl} |.$$

Ce théorème entraîne entre autres l'estimation du rayon spectral

$$\beta\,(A) \leqslant \max_k \sum_{l=1}^{n} | a_{kl} |.$$

L'avènement des ordinateurs de plus en plus puissants fait qu'on préfère souvent de chercher des composantes de la solution qui correspondent à plusieurs premières valeurs propres des problèmes spectraux

$$A\varphi = \lambda\varphi.$$

Cela signifie qu'on procède comme plus haut pour trouver la valeur propre $\lambda_1 = \alpha$ et le vecteur propre associé u_1. Soit A une

matrice symétrique à spectre simple $\{\lambda_n\}$. λ_2 et u_2 s'obtiennent par itération :

$$\psi^{(n+1)} = B \frac{\psi^{(n)}}{\| \psi^{(n)} \|}, \qquad B = \beta(A) E - A,$$

et on suppose que l'approximation nulle $\psi^{(0)} = h$ et les approximations suivantes $\psi^{(n)}$ ne contiennent pas de composantes relatives à u_1. A cet effet, (1.39) doit se dérouler avec orthogonalisation par rapport à u_1 telle que

$$\bar{\psi}^{(n)} = \psi^{(n)} - a_1 u_1, \tag{1.43}$$

a_1 étant une constante choisie à partir de la condition d'orthogonalité de $\bar{\psi}^{(n)}$ et u_1, i.e. à partir de

$$(\bar{\psi}^{(n)}, u_1) = 0. \tag{1.44}$$

Multiplions (1.43) scalairement par u_1, il vient, compte tenu de (1.44),

$$(\bar{\psi}^{(n)}, u_1) = (\psi^{(n)}, u_1) - a_1 (u_1, u_1) = 0.$$

D'où

$$a_1 = \frac{(\psi^{(n)}, u_1)}{(u_1, u_1)}. \tag{1.45}$$

On aboutit à l'algorithme

$$\psi^{(n+1)} = B \frac{\bar{\psi}^{(n)}}{\| \bar{\psi}^{(n)} \|}, \qquad \bar{\psi}^{(n)} = \psi^{(n)} - a_1 u_1. \tag{1.46}$$

Lorsque $n \to \infty$, le processus (1.46) tend vers u_2 et la valeur propre λ_2 s'obtient par la formule

$$\lambda_2 = \lim_{n \to \infty} \frac{\| A\bar{\psi}^{(n)} \|}{\| \bar{\psi}^{(n)} \|}.$$

Les vecteurs suivants et les valeurs propres associées se définissent de façon analogue.

Supposons, en effet, calculés k premiers vecteurs propres $u_1, u_2, \ldots, u_k$. On demande u_{k+1}. Utilisons le processus itératif

$$\psi^{(n+1)} = B \frac{\psi^{(n)}}{\| \psi^{(n)} \|},$$

$$\bar{\psi}^{(n)} = \psi^{(n)} - a_1 u_1 - a_2 u_2 - \ldots - a_k u_k,$$

avec $a_1, a_2, \ldots, a_k$ calculés par la condition d'orthogonalité de $\bar{\psi}^{(n)}$ à tous les $u_1, u_2, \ldots, u_k$. Finalement, on a

$$a_1 (u_1, u_1) = (\psi^{(n)}, u_1),$$
$$a_2 (u_2, u_2) = (\psi^{(n)}, u_2),$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$a_k (u_k, u_k) = (\psi^{(n)}, u_k).$$

D'où

$$a_1 = \frac{(\psi^{(n)}, u_1)}{(u_1, u_1)}, \quad a_2 = \frac{(\psi^{(n)}, u_2)}{(u_2, u_2)}, \ \ldots, \quad a_k = \frac{(\psi^{(n)}, u_k)}{(u_k, u_k)}.$$

Avec n tendant vers l'infini, le processus construit converge vers

$$u_{k+1} = \lim_{n\to\infty} \bar{\psi}^{(n)},$$

et la valeur propre correspondante est calculée par la formule

$$\lambda_{k+1} = \lim_{n\to\infty} \frac{\| A\bar{\psi}^{(n)} \|}{\| \bar{\psi}^{(n)} \|}.$$

Chose à noter, deux orthogonalisations successives doivent être séparées par un nombre relativement important de pas d'itération pour que les erreurs d'arrondi dues aux vecteurs « amortis » ne soient pas trop grandes. Il serait bon d'orthogonaliser les approximations « initiales » $\varphi^{(0)}$.

La méthode décrite illustre une façon de chercher les premières valeurs propres et les vecteurs propres correspondants des matrices symétriques à spectre simple.

Plaçons-nous dans le cas de A non symétrique, i.e. on aborde deux problèmes spectraux :

$$A\varphi = \lambda\varphi, \quad A^*\varphi^* = \lambda\varphi^*. \tag{1.47}$$

Les vecteurs propres et le spectre de (1.47) sont toujours supposés réels et les valeurs propres simples. Les bases $\{\varphi_n\}$ et $\{\varphi_n^*\}$ sont biorthogonales, i.e.

$$(\varphi_n, \varphi_m^*) = 0 \quad \text{si} \quad n \neq m.$$

L'algorithme de génération de premiers vecteurs propres procède comme suit. On trouve

$$u_1 = \lim_{n\to\infty} B \frac{\psi^{(n)}}{\| \psi^{(n)} \|},$$

$$u_1^* = \lim_{n\to\infty} B^* \frac{\psi^{*(n)}}{\| \psi^{*(n)} \|}$$

et la plus petite valeur propre

$$\lambda_1 = \lim_{n\to\infty} \frac{\| A\psi^{(n)} \|}{\| \psi^{(n)} \|} = \lim_{n\to\infty} \frac{\| A^*\psi^{*(n)} \|}{\| \psi^{*(n)} \|},$$

puis on cherche de proche en proche les autres vecteurs propres et les valeurs propres associées. On utilise le processus itératif

$$\psi^{(n+1)} = B \frac{\bar{\psi}^{(n)}}{\| \bar{\psi}^{(n)} \|},$$

$$\psi^{*(n+1)} = B^* \frac{\bar{\psi}^{*(n)}}{\| \bar{\psi}^{*(n)} \|},$$

avec

$$\bar{\psi}^{(n)} = \psi^{(n)} - a_1 u_1 - a_2 u_2 - \ldots - a_k u_k,$$
$$\bar{\psi}^{*(n)} = \psi^{*(n)} - b_1 u_1^* - b_2 u_2^* - \ldots - b_k u_k^*.$$

Les constantes $a_1, \ldots, a_k$ et $b_1, \ldots, b_k$ sont calculées moyennant les conditions d'orthogonalité

$$\begin{aligned} (\bar{\psi}^{(n)}, \ u_i^*) &= 0, \\ (\bar{\psi}^{*(n)}, \ u_i) &= 0, \end{aligned} \qquad i = 1, 2, \ldots, k.$$

Etant donnée la biorthogonalité, il vient

$$\begin{aligned} (\psi^{(n)}, \ u_i^*) - a_i (u_i, \ u_i^*) &= 0, \\ (\psi^{*(n)}, \ u_i) - b_i (u_i, \ u_i^*) &= 0, \end{aligned} \qquad i = 1, 2, \ldots, k.$$

Finalement, on définit les constantes a_i, b_i par les égalités

$$a_i = \frac{(\psi^{(n)}, \ u_i^*)}{(u_i, \ u_i^*)}, \quad b_i = \frac{(\psi^{*(n)}, \ u_i)}{(u_i, \ u_i^*)}, \qquad i = 1, 2, \ldots, k.$$

La transformation QR inspire aujourd'hui une classe d'algorithmes plus efficaces pour la résolution complète des problèmes de valeurs propres. Grâce à Frencis [VIII], Koublanovskaïa [VIII ; 8], Wilkinson [VIII], pour ne citer que ces auteurs, on a vu apparaître plusieurs programmes et procédures de première qualité pour la recherche des vecteurs et valeurs propres des problèmes d'algèbre linéaire, qui s'appuient sur l'algorithme QR. Ces programmes et ces procédures concurrencent avec succès les schémas itératifs et ils s'imposeront visiblement dans des classes étendues de problèmes, y compris pour les matrices d'ordre élevé. Dans le dernier cas, l'efficacité de ces algorithmes dépend essentiellement de la puissance des ordinateurs. Or, celle-ci augmente sans cesse.

Selon l'auteur, la résolution complète des problèmes de valeurs propres va déterminer, dans le cadre de l'évolution générale de l'Analyse numérique, le renouveau de la méthode classique de Fourier en ce qui concerne les problèmes de la physique mathématique qui se ramènent à des problèmes d'algèbre linéaire.

En effet, soit le problème d'algèbre linéaire

$$A\varphi = f, \tag{1.48}$$

dont on demande les solutions avec le même opérateur A et f différents. Connaissant les solutions des problèmes spectraux

$$Aw = \lambda w,$$
$$A^* w^* = \lambda w^*$$

sous forme de systèmes complets $\{w_n\}$ et $\{w_n^*\}$, la solution de (1.48) et le second membre f doivent être cherchés comme sommes de Fourier

$$\varphi = \sum_{n=1}^{m} \varphi_n w_n, \qquad f = \sum_{n=1}^{m} f_n w_n, \tag{1.49}$$

où

$$f_n = (f,\ w_n^*).$$

Portons (1.49) dans (1.48) et multiplions scalairement par w_n^*, il vient le système algébrique

$$\lambda_n \varphi_n = f_n, \qquad n = 1,\ 2,\ \ldots,\ m,$$

qui entraîne pour $\lambda_n \neq 0$

$$\varphi_n = \frac{f_n}{\lambda_n}.$$

Ainsi, la solution du problème (1.48) s'écrit

$$\varphi = \sum_{n=1}^{m} \frac{f_n}{\lambda_n} w_n. \tag{1.50}$$

Lorsque les vecteurs propres connus du système sont en nombre restreint, disons $k < m$, la méthode de Fourier permet de chercher cette solution sous la forme

$$\varphi = \sum_{n=1}^{k} \frac{f_n}{\lambda_n} w_n + \xi. \tag{1.51}$$

Ici ξ vérifie l'équation

$$A\xi = \eta, \tag{1.52}$$

avec

$$\eta = f - \sum_{n=1}^{k} f_n w_n.$$

Il paraît de prime abord que (1.48) est ramené au problème analogue (1.52). Formellement, il en est ainsi. Mais si la matrice A est d'ordre élevé au point que les deux problèmes en question ne se prêtent qu'à des procédés itératifs (voir Chapitre 4), il s'avère que s'agissant de (1.52) ceux-ci sont beaucoup plus commodes en calcul automatique. Ce fait détermine l'intérêt de l'algorithme décrit destiné à fournir

les premiers vecteurs propres du problème spectral. Ces idées nous seront utiles plus d'une fois lorsque nous aurons à former des procédures de calcul.

1.1.3. Valeurs propres et fonctions propres du laplacien

Soit

$$A = -\Delta, \tag{1.53}$$

$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ étant le laplacien. On définit l'opérateur $A = -\Delta$ sur un ensemble Φ dont les éléments satisfont aux conditions suivantes. *Primo*,

$$\varphi = 0 \quad \text{sur} \quad \partial D, \tag{1.54}$$

avec ∂D la frontière de D. On admet pour simplifier que $D = \{(x, y),\ 0 < x < 1,\ 0 < y < 1\}$.

Secundo, toute $\varphi(x)$ est continue dans le domaine fermé $D + \partial D$ ainsi que ses dérivées d'ordre 1 et 2.

Tertio, les fonctions φ forment un sous-espace Φ dans l'espace hilbertien $L_2(D)$ avec le produit scalaire

$$(a, b) = \int_D ab\, dD, \tag{1.55}$$

où $a \in L_2(D)$, $b \in L_2(D)$, et la norme

$$\|\varphi\| = \sqrt{(\varphi, \varphi)}. \tag{1.56}$$

Etant données ces hypothèses, on montre la symétrie de l'opérateur A. Considérons une fonction $\varphi^* \in L_2(D)$ et la fonctionnelle

$$(A\varphi, \varphi^*) = -\int_D \varphi^* \Delta\varphi\, dD \tag{1.57}$$

(les conditions imposées à φ garantissent la borne de $(A\varphi, \varphi^*)$ quels que soient φ^*). On suppose que φ^* est suffisamment régulière et on obtient moyennant la seconde formule de Green

$$(A\varphi, \varphi^*) = -\int_{\partial D} \left(\varphi^* \frac{\partial\varphi}{\partial n} - \varphi \frac{\partial\varphi^*}{\partial n}\right) ds - \int_D \varphi \Delta\varphi^*\, dD, \tag{1.58}$$

n étant la normale extérieure à D. Si φ^* remplit la condition aux limites

$$\varphi^* = 0 \text{ sur } \partial D, \tag{1.59}$$

on a, compte tenu des conditions (1.54) et (1.59),

$$(A\varphi, \varphi^*) = -\int_D \varphi \Delta\varphi^*\, dD = (\varphi, A\varphi^*), \tag{1.60}$$

i.e. $A \equiv A^*$ et A est bien symétrique. Il ressort de l'analyse des transformations effectuées que φ^* doit posséder des dérivées continues *). Conclusion: A est auto-adjoint sur Φ.

Voyons quelle est la définitivité de A. Considérons la fonctionnelle

$$(A\varphi,\ \varphi) = -\int_D \varphi\,\Delta\varphi\,dD. \tag{1.61}$$

On a par la première formule de Green:

$$(A\varphi,\ \varphi) = -\int_{\partial D} \varphi\frac{\partial\varphi}{\partial n}\,ds + \int_D \left[\left(\frac{\partial\varphi}{\partial x}\right)^2 + \left(\frac{\partial\varphi}{\partial y}\right)^2\right] dD. \tag{1.62}$$

Comme φ remplit la condition (1.54), on a

$$(A\varphi,\ \varphi) = \int_D \left[\left(\frac{\partial\varphi}{\partial x}\right)^2 + \left(\frac{\partial\varphi}{\partial y}\right)^2\right] dD > 0 \tag{1.63}$$

pour toute $\varphi \in \Phi$ non identiquement nulle.

Considérons dans ce contexte un problème de valeurs propres. Il est connu (Courant [II], Sobolev [I; 6]) que le système orthonormé de fonctions propres du problème

$$Au = \lambda u \text{ dans } D \tag{1.64}$$

est complet si

$$u = 0 \quad \text{sur} \quad \partial D. \tag{1.65}$$

Il est de la forme

$$u_{mp} = 2 \sin m\pi x \sin p\pi y, \tag{1.66}$$

où $m = 1, 2, \ldots$ et $p = 1, 2, \ldots$, les valeurs propres de l'opérateur A étant

$$\lambda_{mp}(A) = (m^2 + p^2)\,\pi^2 > 0. \tag{1.67}$$

Il en résulte que

$$2\pi^2 \leqslant \lambda_{mp}(A) < \infty.$$

Ainsi,

$$\alpha(A) = 2\pi^2, \qquad \beta(A) = \infty \tag{1.68}$$

et $(A\varphi,\ \varphi) \geqslant 2\pi^2(\varphi,\ \varphi)$. A est donc défini positif et non borné. Vu la complétude du système de fonctions propres $\{u_{mp}\}$, toute fonction de Φ se développe en série de Fourier

$$\varphi(x,\ y) = \sum_m \sum_p \varphi_{mp} u_{mp}(x,\ y) = \sum_i \varphi_i u_i(x,\ y). \tag{1.69}$$

*) Cette analyse a lieu sous des contraintes très sévères sur la solution. On montre que la formule de Green et, partant, tous nos raisonnements restent en vigueur pour toute φ de l'espace de Sobolev $\mathring{W}_2^2$.

Le système $\{u_{mp}\}$ est orthonormé, si bien que

$$\varphi_i = (\varphi, u_i), \tag{1.70}$$

avec i le nouvel indice repérant le numéro des termes.

Nous avons considéré le problème de Dirichlet pour le laplacien dans un carré. S'agissant des domaines plus compliqués et des opérateurs à coefficients variables, les valeurs propres ne s'obtiennent pas sous forme explicite et on localise le spectre par des techniques spéciales.

1.1.4. Valeurs propres et vecteurs propres du laplacien aux différences finies

Soit un ensemble de points (x_k, y_l), où $x_k = k/n$, $y_l = l/n$, n étant un entier naturel donné et k et l des entiers naturels quelconques. Cet ensemble ponctuel constitue un *réseau*, ses points sont les *nœuds du réseau* et la quantité $h = 1/n$ est le *pas*. Introduisons la notation $D_h = \{(x_k, y_l),\ 1 \leqslant k < n,\ 1 \leqslant l < n\}$ et appelons ∂D_h le réseau basé sur les nœuds dont l'une des coordonnées est soit 0, soit 1. Les fonctions définies sur un réseau sont des *fonctions discrètes*. Φ_h particularisera l'ensemble des fonctions discrètes φ^h définies dans D_h. On associe à chaque $\varphi \in \Phi$ une fonction discrète $(\varphi)_h$ selon la loi suivante : la valeur de $(\varphi)_h$ au nœud (x_k, y_l) est $\varphi(x_k, y_l)$. La correspondance indiquée est un opérateur linéaire de Φ dans Φ_h (sous-espace des fonctions discrètes sur D_h) qu'on appelle *projection* de φ sur le réseau.

Soit A un opérateur linéaire défini sur $\varphi \in \Phi$, auquel cas on projette $\psi = A\varphi$ sur le réseau à condition de poser $(\psi)_h = (A\varphi)_h$. La correspondance entre $(\varphi)_h$ et $(A\varphi)_h$ est un opérateur linéaire défini sur les fonctions discrètes. Il constitue une projection de A sur le réseau et se note $(A)_h$. Cette projection sur un domaine discrétisé conduit à des équations aux différences finies. (Les méthodes de construction correspondantes, l'approximation, la stabilité numérique et la convergence de la solution du problème approché vers la solution du problème exact seront examinées dans la suite.)

Soit φ^h une fonction discrète de composantes $\varphi^h_{k,l}$ et Δ^h le laplacien aux différences associé à un réseau régulier de pas h défini par l'égalité

$$(\Delta^h \varphi^h)_{k,l} = \frac{\varphi^h_{k+1,l} + \varphi^h_{k-1,l} + \varphi^h_{k,l+1} + \varphi^h_{k,l-1} - 4\varphi^h_{k,l}}{h^2}. \tag{1.71}$$

Supposons $\varphi^h \in \Phi_h$ nulle sur la frontière du domaine discrétisé, i.e.

$$(\varphi^h)_{k,l} = 0 \quad \text{sur} \quad \partial D_h, \tag{1.72}$$

introduisons les opérateurs aux différences

$$(\Delta_x \varphi^h)_{k,l} = \frac{1}{h}(\varphi^h_{k+1,l} - \varphi^h_{k,l}), \quad (\nabla_x \varphi^h)_{k,l} = \frac{1}{h}(\varphi^h_{k,l} - \varphi^h_{k-1,l})$$

et de même

$$(\Delta_y \varphi^h)_{k,\,l} = \frac{1}{h}(\varphi^h_{k,\,l+1} - \varphi^h_{k,\,l}), \quad (\nabla_y \varphi^h)_{k,\,l} = \frac{1}{h}(\varphi^h_{k,\,l} - \varphi^h_{k,\,l-1}).$$

Soient A^h, A_x et A_y des opérateurs aux différences définis par

$$\begin{gathered} A^h = A_x + A_y, \\ A_x = -\Delta_x \nabla_x, \quad A_y = -\Delta_y \nabla_y. \end{gathered} \tag{1.73}$$

On a

$$-\Delta^h = A_x + A_y = A^h.$$

L'ensemble des nœuds tels que ou bien $k = 0$, n, ou bien $l = 0$, n forme ∂D_h (on rappelle que selon (1.72) φ^h s'annule en ces points nodaux). Dans la suite, on omet h indiçant les fonctions discrètes φ et φ^*.

Soit le produit scalaire

$$(a,\, b) = h^2 \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} a_{k,\,l} b_{k,\,l}$$

et la norme

$$\| \varphi \| = \sqrt{(\varphi,\, \varphi)}.$$

Formons la fonctionnelle

$$(A^h \varphi,\, \varphi^*) = -h^2 \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} [(\Delta_x \nabla_x \varphi)_{k,\,l} + (\Delta_y \nabla_y \varphi)_{k,\,l}]\, \varphi^*_{k,\,l}.$$

On a les identités suivantes analogues aux formules de Green (Ladyjenskaïa [II; 5], Samarski [III; 12]):

$$\begin{gathered} -\sum_{k=1}^{n-1} (\Delta_x \nabla_x \varphi)_{k,\,l} \varphi^*_{k,\,l} = \sum_{k=1}^{n} (\nabla_x \varphi)_{k,\,l} (\nabla_x \varphi^*)_{k,\,l}, \\ -\sum_{k=1}^{n-1} (\Delta_x \nabla_x \varphi)_{k,\,l} \varphi^*_{k,\,l} = -\sum_{k=1}^{n-1} (\Delta_x \nabla_x \varphi^*)_{k,\,l} \varphi_{k,\,l}. \end{gathered} \tag{1.74}$$

Les formules (1.74) ne sont justes que pour $\varphi \in \Phi_h$ avec (1.72) et $\varphi^* \in \Phi^*_h$ telles que

$$\varphi^*_{k,\,l} = 0 \quad \text{sur} \quad \partial D_h. \tag{1.75}$$

Des égalités analogues ont lieu pour les sommes par rapport à l. On obtient, à l'aide de la seconde relation (1.74),

$$(A^h \varphi,\, \varphi^*) = (\varphi,\, A^h \varphi^*),$$

d'où

$$A^h = (A^h)^* \quad \text{et} \quad \Phi(A^h) = \Phi^*((A^h)^*).$$

Considérons la fonctionnelle

$$(A^h\varphi,\ \varphi) = -h^2 \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} [(\Delta_x \nabla_x \varphi)_{k,\,l} + (\Delta_y \nabla_y \varphi)_{k,\,l}]\, \varphi_{k,\,l}.$$

La première relation (1.74) donne pour k et l :

$$(A^h\varphi,\ \varphi) = h^2 \sum_{k=1}^{n} \sum_{l=1}^{n} [((\nabla_x \varphi)_{k,\,l})^2 + ((\nabla_y \varphi)_{k,\,l})^2],$$

d'où

$$(A^h\varphi,\ \varphi) > 0$$

si φ n'est pas le vecteur nul.

Soit enfin le problème spectral

$$\begin{aligned} A^h u &= \lambda u \quad \text{dans} \quad D_h, \\ u &= 0 \quad \text{sur} \quad \partial D_h. \end{aligned} \tag{1.76}$$

Les vecteurs propres orthonormés associés ont pour composantes

$$u_{mp}^{kl} = 2 \sin m\pi kh \sin p\pi lh, \tag{1.77}$$

$$m = 1,\ 2,\ \ldots,\ n-1;\quad p = 1,\ 2,\ \ldots,\ n-1.$$

On rappelle que $(u_{m_1p_1},\ u_{m_2p_2}) = h^2 \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} u_{m_1p_1}^{k,\,l} u_{m_2p_2}^{k,\,l}$. Dans (1.77) k et l indicent les composantes et m et p les numéros des vecteurs propres qu'on range comme suit :

$$u_{mp} = u_i, \qquad i = 1,\ 2,\ \ldots$$

Etant données les relations évidentes

$$-(\Delta_x \nabla_x u_i)_{k,\,l} = 2 \left(\frac{4}{h^2} \sin^2 \frac{m\pi h}{2} \sin m\pi kh \right) \sin p\pi lh,$$

$$-(\Delta_y \nabla_y u_i)_{k,\,l} = 2 \sin m\pi kh \left(\frac{4}{h^2} \sin^2 \frac{p\pi h}{2} \sin p\pi lh \right),$$

les valeurs propres s'écrivent

$$\lambda_{mp}(A^h) = \frac{4}{h^2} \left(\sin^2 \frac{m\pi h}{2} + \sin^2 \frac{p\pi h}{2} \right). \tag{1.78}$$

m et p varient entre 1 et $n-1$. Donc $h = \frac{1}{n} \leqslant mh \leqslant (n-1)\,h = 1 - h$ et $h \leqslant ph \leqslant 1 - h$, si bien que

$$\frac{8}{h^2} \sin^2 \frac{\pi h}{2} \leqslant \lambda_i(A^h) \leqslant \frac{8}{h^2} \cos^2 \frac{\pi h}{2}.$$

Ici $\lambda_i(A^h)$ représentent $\lambda_{mp}(A^h)$ ordonnés de façon adéquate. Puisque $\frac{\pi h}{2}$ est de règle $\ll 1$, on a les égalités approchées

$$\sin^2 \frac{\pi h}{2} = \frac{\pi^2 h^2}{4} - O(h^4), \quad \cos^2 \frac{\pi h}{2} = 1 - O(h^2),$$

donc

$$\alpha(A^h) \leqslant \lambda_i \leqslant \beta(A^h), \tag{1.79}$$

où

$$\alpha(A^h) = \frac{1}{\| (A^h)^{-1} \|} \approx 2\pi^2, \quad \beta(A^h) = \| A^h \| \approx \frac{8}{h^2}. \tag{1.80}$$

Quel que soit $h \neq 0$, l'opérateur A^h est borné, tandis que l'opérateur différentiel ne l'a été pas. Ainsi, la réduction effectuée « a amélioré » les qualités spectrales de l'opérateur.

La base de vecteurs propres (1.77) permet de développer φ en série. On a

$$\varphi = \sum_i \varphi_i u_i, \tag{1.81}$$

où

$$\varphi_i = (\varphi, u_i). \tag{1.82}$$

1.2. Approximation

Soit un problème opératoriel de la physique mathématique

$$\begin{aligned} A\varphi &= f \text{ dans } D, \\ a\varphi &= g \text{ sur } \partial D, \end{aligned} \tag{2.1}$$

avec A un opérateur linéaire, $\varphi \in \Phi$ et $f \in F$. Ici Φ et F sont deux espaces de Hilbert d'éléments définis respectivement dans $D + \partial D$ et D, a est l'opérateur linéaire de condition aux limites, $g \in G$, G étant un espace hilbertien de fonctions définies sur ∂D.

Soit, dans l'espace n-dimensionnel de fonctions discrètes, l'équation

$$\begin{aligned} A^h \varphi^h &= f^h \quad \text{dans} \quad D_h, \\ a^h \varphi^h &= g^h \quad \text{sur} \quad \partial D_h, \end{aligned} \tag{2.2}$$

avec A^h un opérateur linéaire dépendant du pas h, $\varphi^h \in \Phi_h$, $f^h \in F_h$, Φ_h et F_h deux espaces de fonctions discrètes. Ici D_h est l'ensemble des nœuds intérieurs de D, ∂D_h l'ensemble des nœuds frontières, a^h un opérateur linéaire, $g^h \in G_h$, G_h étant l'espace euclidien des vecteurs définis sur ∂D_h.

Prenons sur F_h, G_h, Φ_h les normes respectives $\|\cdot\|_{F_h}$, $\|\cdot\|_{G_h}$, $\|\cdot\|_{\Phi_h}$. Soit $(\cdot)_h$ l'opérateur linéaire qui associe à $\varphi \in \Phi$ un élément $(\varphi)_h \in \Phi_h$ tel que $\lim\limits_{h \to 0} \| (\varphi)_h \|_{\Phi_h} = \| \varphi \|_{\Phi}$.

Nous dirons que le problème (2.2) *approche* le problème (2.1) *à l'ordre n* par rapport à la solution φ s'il existe des constantes positi-

ves $\overline{h}$, M_1, M_2 telles qu'on ait pour tous les $h < \overline{h}$

$$\begin{aligned} \| A^h(\varphi)_h - f^h \|_{F_h} &\leqslant M_1 h^{n_1}, \\ \| a^h(\varphi)_h - g^h \|_{G_h} &\leqslant M_2 h^{n_2} \end{aligned} \tag{2.3}$$

et $n = \min(n_1, n_2)$.

Lorsque la solution de (2.1) est suffisamment régulière, on cherche l'ordre d'approximation moyennant la norme naturelle de l'espace des fonctions continues dérivables. On utilise en l'occurrence le développement taylorien de la solution et des autres fonctions du problème.

Pour plus de détails le lecteur est prié de s'adresser à Godounov, Riabenki [III; 33, 34], Richtmyer [III], Kantorovich, Akilov [I], Samarski [III; 12].

Dans la suite nous supposerons que le problème (2.1) a été ramené à la forme (2.2) et qu'on a de plus éliminé la solution aux points frontières de $D_h + \partial D_h$ moyennant la condition aux limites du dernier problème. On a en définitive

$$\widetilde{A}^h \widetilde{\varphi}^h = \widetilde{f}^h, \tag{2.4}$$

et la solution aux points frontières s'obtient de (2.2) après avoir résolu (2.4).

On utilise selon le cas soit l'écriture (2.4), soit (2.2).

Ainsi, le problème à argument continu (2.1) se trouve ramené au problème d'algèbre linéaire (2.4). Il reste à résoudre le système d'équations algébriques.

EXEMPLE. Soit le problème

$$-\Delta\varphi = f \quad \text{dans} \quad D, \quad \varphi = 0 \quad \text{sur} \quad \partial D. \tag{2.5}$$

On suppose que le domaine de définition D est le carré $\{0 < x < 1, 0 < y < 1\}$ et que f est une fonction régulière. Recouvrons le carré $\overline{D} = D + \partial D$ par un réseau régulier de pas h en x et y. Les nœuds seront indicés par k et l: k $(0 \leqslant k \leqslant n)$ indique la discrétisation par rapport à x et l $(0 \leqslant l \leqslant n)$ joue le même rôle pour y. Considérons les approximations

$$\varphi_{xx} \to \Delta_x \nabla_x (\varphi)_h, \qquad \varphi_{yy} \to \Delta_y \nabla_y (\varphi)_h,$$

avec Δ_x, Δ_y, ∇_x, ∇_y les opérateurs aux différences définis au point 1.1.4. On approche le problème (2.5) par

$$\begin{aligned} -[\Delta_x \nabla_x \varphi^h + \Delta_y \nabla_y \varphi^h] &= f^h \quad \text{dans} \quad D_h, \\ \varphi^h_{k,l} &= 0 \quad \text{sur} \quad \partial D_h, \end{aligned} \tag{2.6}$$

∂D_h étant l'ensemble des nœuds frontières. Il est déjà apparu que le problème (2.6) peut se ramener à

$$\begin{aligned} -\Delta^h \varphi^h &= f^h \quad \text{dans} \quad D_h, \\ \varphi^h &= 0 \quad \text{sur} \quad \partial D_h, \end{aligned} \tag{2.7}$$

avec φ^h et f^h deux vecteurs de composantes respectives $\varphi^h_{k,\,l}$ et $f^h_{k,\,l}$ et

$$(\Delta^h \varphi^h)_{k,\,l} = \frac{1}{h^2}(\varphi^h_{k+1,\,l} + \varphi^h_{k-1,\,l} + \varphi^h_{k,\,l+1} + \varphi^h_{k,\,l-1} - 4\varphi^h_{k,\,l}),$$

$$f^h_{k,\,l} = \frac{1}{h^2} \int_{x_{k-1/2}}^{x_{k+1/2}} \int_{y_{l-1/2}}^{y_{l+1/2}} f\,dx\,dy$$

$$\left(x_{k\pm 1/2} = x_k \pm \frac{h}{2}, \quad y_{l\pm 1/2} = y_l \pm \frac{h}{2}\right).$$

Noterons Φ_h l'espace des solutions définies dans $D_h + \partial D_h = \{(x_k, y_l);\ 0 \leqslant k \leqslant n,\ 0 \leqslant l \leqslant n\}$. Le vecteur $f^h \in F_h$ a pour domaine $D_h = \{(x_k, y_l);\ 1 \leqslant k \leqslant n-1,\ 1 \leqslant l \leqslant n-1\}$. Développons la solution en série de Taylor au voisinage du point (x_k, y_l) et admettons que les dérivées par rapport à x et y jusqu'à l'ordre 4 inclus sont continues, il vient

$$\varphi(\bar{x}, \bar{y}) = \sum_{n=0}^{3} \frac{1}{n!} \left\{\left[(\bar{x} - x_k)\frac{\partial}{\partial x} + (\bar{y} - y_l)\frac{\partial}{\partial y}\right]^n \varphi\right\}_{k,\,l} +$$

$$+ \frac{1}{4!} \left\{\left[(\bar{x} - x_k)\frac{\partial}{\partial x} + (\bar{y} - y_l)\frac{\partial}{\partial y}\right]^4 \varphi\right\}_{k+\theta_1,\,l+\theta_2},$$

avec $(\bar{x}, \bar{y})$ un point quelconque du domaine

$$\{x_{k-1} \leqslant x \leqslant x_{k+1},\quad y_{l-1} \leqslant y \leqslant y_{l+1}\},\quad |\theta_1|,\ |\theta_2| < 1,$$

et

$$x_{k+\theta_1} = x_k + \theta_1 h, \qquad y_{l+\theta_2} = y_l + \theta_2 h.$$

La fonction $f(x, y)$ admet un développement analogue.

Munissons F_h de la norme $\|f^h\|_{F_h} = \max\limits_{k,\,l} |f^h_{k,\,l}|$ et procédons de façon analogue avec G_h. Prenons pour $(\varphi)_h$ le vecteur formé de valeurs de φ aux nœuds correspondants. Les développements ci-dessus de φ et f donnent

$$\|-\Delta^h (\varphi)_h - f^h\|_{F_h} \leqslant M_1 h^2, \tag{2.8}$$

avec $M_1 = \frac{1}{6} \max\limits_{\overline{D}} (|\varphi^{\mathrm{IV}}_x|,\ |\varphi^{\mathrm{IV}}_y|)$. Dans ce cas, les conditions aux limites sont approchées exactement.

Il découle de (2.8) que le problème (2.7) approche (2.5) à l'ordre 2 par rapport aux solutions de (2.5) possédant des dérivées quatrièmes bornées.

Dans les cas considérés, on s'est intéressé aux seules variables d'espace. Soit le problème d'approcher l'équation d'évolution *)

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + A\varphi = f \quad \text{dans} \quad D \times D_t, \\ &a\varphi = g \quad \text{sur} \quad \partial D \times D_t, \\ &\varphi = \varphi^0 \quad \text{dans} \quad D \quad \text{pour} \quad t = 0. \end{aligned} \tag{2.9}$$

L'approximation s'effectue en deux étapes. On approche d'abord dans $(D_h + \partial D_h) \times D_t$ par rapport aux variables d'espace et on obtient une équation qui est différentielle en t et discrétisée en les variables spatiales.

Dans ce problème hybride, on élimine des fois facilement les solutions aux points frontières de $(D_h + \partial D_h) \times D_\tau$ moyennant les conditions aux limites discrétisées. Une fois cette élimination faite, on aboutit à une équation d'évolution de la forme

$$\frac{d\varphi^h}{dt} + \Lambda\varphi^h = f^h, \tag{2.10}$$

où Λ, f^h et φ^h dépendent de t. Nous omettrons l'indice h dans (2.10) que nous assimilerons à un analogue discrétisé par rapport aux variables d'espace du problème initial.

L'équation (2.10) constitue évidemment un système d'équations différentielles ordinaires pour les coordonnées de φ^h. Soit le problème de Cauchy

$$\begin{aligned} &\frac{d\varphi}{dt} + \Lambda\varphi = f, \\ &\varphi = g \quad \text{pour} \quad t = 0, \end{aligned} \tag{2.11}$$

l'opérateur Λ étant supposé indépendant de t. Comment approcher (2.11) par rapport au temps ? On utilise le plus volontiers les schémas aux différences d'ordre 1 et 2 en t.

Considérons d'abord le schéma explicite simple qui approche à l'ordre 1 sur le réseau D_τ

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + \Lambda\varphi^j = f^j, \qquad \varphi^0 = g, \tag{2.12}$$

où $\tau = t_{j+1} - t_j$, f^j est une projection de f. Pour simplifier les notations nous posons $f^j = f(t_j)$.

*) Une équation est dite *d'évolution* si elle est résoluble explicitement par rapport à la dérivée première en t et si les dérivées par rapport au temps ne figurent pas dans A.

S'agissant d'un schéma implicite simple, on a

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\varphi^{j+1}=f^j, \quad \varphi^0=g, \tag{2.13}$$

et on prend f^j égal à $f(t_{j+1})$. Les deux schémas approchent à l'ordre 1 par rapport au temps. On le vérifie sans peine en développant en série de Taylor par rapport à t sous l'hypothèse que la solution admet par exemple des dérivées secondes bornées (par rapport à t).

Résolvons (2.12) et (2.13) par rapport à l'inconnue, il vient la relation récurrentielle

$$\varphi^{j+1}=T\varphi^j+\tau Sf^j, \tag{2.14}$$

T et S désignant respectivement l'*opérateur de transition* et l'*opérateur de source* définis comme suit : $T=E-\tau\Lambda$, $S=E$ pour (2.12) ; $T=(E+\tau\Lambda)^{-1}$, $S=T$ pour (2.13).

Nous dirons que les schémas du type (2.14) pour les équations d'évolution sont *à deux niveaux*.

Dans les applications, on recourt souvent au schéma de Crank-Nicholson qui réalise une approximation d'ordre 2 :

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\frac{\varphi^{j+1}+\varphi^j}{2}=f^j, \quad \varphi^0=g, \tag{2.15}$$

où $f^j=f(t_{j+1/2})$. Ce schéma se met sous forme (2.14) si

$$T=\left(E+\frac{\tau}{2}\Lambda\right)^{-1}\left(E-\frac{\tau}{2}\Lambda\right),$$

$$S=\left(E+\frac{\tau}{2}\Lambda\right)^{-1}.$$

Il y a des fois intérêt à écrire les équations aux différences (2.12), (2.13) et (2.15) comme système de deux équations dont l'une approche l'équation même dans $D_{h\tau}$ et l'autre la condition aux limites sur $\partial D_{h\tau}$. L'analogue aux différences du problème (2.9) se met sous la forme

$$\begin{aligned} L^{h\tau}\varphi^{h\tau}&=f^{h\tau} \quad \text{dans} \quad D_{h\tau},\\ l^{h\tau}\varphi^{h\tau}&=g^{h\tau} \quad \text{sur} \quad \partial D_{h\tau}, \end{aligned} \tag{2.16}$$

où

$$\begin{aligned} &D_{h\tau}=D_h\times D_\tau, \qquad \partial D_{h\tau}=D_h\times\{0\}\cup\partial D_h\times D_\tau,\\ &\|L^{h\tau}(\varphi)_{h\tau}-f^{h\tau}\|_{F_{h\tau}}\leqslant M_1h^n+N_1\tau^p,\\ &\|l^{h\tau}(\varphi)_{h\tau}-g^{h\tau}\|_{G_{h\tau}}\leqslant M_2h^n+N_2\tau^p. \end{aligned} \tag{2.17}$$

A l'instar de (2.3), $(\cdot)_{h\tau}$ est l'opérateur de projection sur l'espace discrétisé correspondant.

L'équation aux différences sous forme canonique (2.14) se récrit

$$L\varphi = f \tag{2.18}$$

si l'on introduit des fonctions vectorielles et de nouveaux opérateurs dans $D_h \times D_\tau$, D_τ étant l'ensemble $\{t_j\}$.

Ainsi, l'équation d'évolution a été réduite, compte tenu des conditions aux limites et des données initiales, au problème d'algèbre linéaire (2.18). On utilise selon le cas soit le réseau D_h, soit $D_h \times D_\tau$. Dans un cas particulier, on ramène à (2.18) un problème aux limites du type elliptique, une équation intégrale, et ainsi de suite. La condition d'approximation s'écrit de nouveau sous forme (2.17) avec le seul paramètre de discrétisation h, valeur maximale de l'ensemble $\{\Delta x_j\}$ des pas d'espace.

EXEMPLE. Considérons le problème

$$\begin{aligned} A\varphi \equiv \frac{\partial \varphi}{\partial t} - \Delta\varphi = f \quad &\text{dans} \quad D \times D_t, \\ \varphi = 0 \quad &\text{sur} \quad \partial D \times D_t, \\ \varphi = g \quad &\text{dans} \quad D \quad \text{pour} \quad t = 0. \end{aligned} \tag{2.19}$$

La solution sera supposée définie dans $(D + \partial D) \times D_t$, D étant encore un carré et $D_t = \{0 \leqslant t \leqslant T\}$. Passons de D à D_h, de ∂D à ∂D_h et de D_t à D_τ. Soit D_τ un ensemble des points t_j et $t_{j+1} - t_j = \tau$. Le problème (2.19) peut être approché par

$$\begin{aligned} A^{h\tau}\varphi^j = f^j \quad &\text{dans} \quad D_h \times D_\tau, \\ \varphi^j = 0 \quad &\text{sur} \quad \partial D_h \times D_\tau, \\ \varphi^0 = g \quad &\text{dans} \quad D_h \times \{0\}. \end{aligned} \tag{2.20}$$

Considérons le schéma d'approximation explicite simple

$$(A^{h\tau}\varphi)^j_{k,l} \equiv \frac{\varphi^{j+1}_{k,l} - \varphi^j_{k,l}}{\tau} - \Delta^h \varphi^j_{k,l}, \tag{2.21}$$

$$f^j_{k,l} = \frac{1}{h^2} \int_{x_{k-1/2}}^{x_{k+1/2}} \int_{y_{l-1/2}}^{y_{l+1/2}} f(x, y, t_j)\, dx\, dy, \tag{2.22}$$

$$g_{k,l} = \frac{1}{h^2} \int_{x_{k-1/2}}^{x_{k+1/2}} \int_{y_{l-1/2}}^{y_{l+1/2}} g(x, y)\, dx\, dy, \tag{2.23}$$

auquel cas

$$\varphi^{j+1}_{k,l} = \varphi^j_{k,l} + \tau\Delta^h \varphi^j_{k,l} + \tau f^j_{k,l} \quad \text{dans} \quad D_h \times D_\tau. \tag{2.24}$$

De plus,

$$\begin{aligned} \varphi^j_{k,l} = 0 \quad &\text{sur} \quad \partial D_h \times D_\tau, \\ \varphi^0_{k,l} = g_{k,l} \quad &\text{dans} \quad D_h \times \{0\}. \end{aligned} \tag{2.25}$$

La relation de récurrence (2.24) se récrit

$$\varphi_{k,l}^{j+1} = T\varphi_{k,l}^{j} + \tau f_{k,l}^{j}, \tag{2.26}$$

avec $T = E + \tau\Delta^h = E - \tau(A_1 + A_2)$ l'opérateur de transition, A_i ($A_1 = A_x$, $A_2 = A_y$) étant définis par (1.73). Evaluons la norme de T et cherchons à cet effet la plus grande valeur propre de T:

$$\begin{aligned} Tu &= \lambda(T)\, u \text{ dans } D_h, \\ u &= 0 \text{ sur } \partial D_h. \end{aligned} \tag{2.27}$$

On a l'égalité évidente

$$\lambda_n(T) = 1 + \tau\lambda_n(\Delta^h).$$

La norme cherchée est donc de la forme

$$\|T\| = \max\left\{\left|1 - \frac{8\tau}{h^2}\cos^2\frac{\pi h}{2}\right|,\ \left|1 - \frac{8\tau}{h^2}\sin^2\frac{\pi h}{2}\right|\right\}, \tag{2.28}$$

et si $\frac{\tau}{h^2} < \frac{1}{4}$, alors $\|T\| < 1$.

A part l'approximation explicite d'ordre 1 en τ, on considère une approximation *implicite* qui est d'ordre 1 en τ et d'ordre 2 en h. L'expression (2.21) fait alors place à

$$(A^{h\tau}\varphi)_{k,l}^{j} \equiv \frac{\varphi_{k,l}^{j+1} - \varphi_{k,l}^{j}}{\tau} - \Delta^h\varphi_{k,l}^{j+1}. \tag{2.29}$$

Les quantités $f_{k,l}^{j}$ et $g_{k,l}$ sont données par les formules (2.22) et (2.23). Dans ce cas, (2.20) ne s'explicite plus et on a l'équation opératorielle

$$(E - \tau\Delta^h)\,\varphi_{k,l}^{j+1} = \varphi_{k,l}^{j} + \tau f_{k,l}^{j} \quad \text{dans} \quad D_h \times D_\tau \tag{2.30}$$

qu'on résout sous les conditions

$$\begin{aligned} \varphi_{k,l}^{j} &= 0 \qquad \text{sur} \quad \partial D_h \times D_\tau, \\ \varphi_{k,l}^{0} &= g_{k,l} \quad \text{dans} \quad D_h \times \{0\}. \end{aligned} \tag{2.31}$$

Récrivons (2.30):

$$\varphi_{k,l}^{j+1} = T(\varphi_{k,l}^{j} + \tau f_{k,l}^{j}), \tag{2.32}$$

avec

$$T = (E - \tau\Delta^h)^{-1}.$$

La norme de l'opérateur T vaut alors

$$\|T\| = \max\left\{\frac{1}{1 + \frac{8\tau}{h^2}\cos^2\frac{\pi h}{2}},\ \frac{1}{1 + \frac{8\tau}{h^2}\sin^2\frac{\pi h}{2}}\right\}; \tag{2.33}$$

donc $\|T\| < 1$ quels que soient τ et h.

Utilisons le schéma de Crank-Nicholson et définissons les opérateurs et les fonctions du problème (2.20) par les égalités

$$(A^{h\tau}\varphi)^j_{k,l} \equiv \frac{\varphi^{j+1}_{k,l}-\varphi^j_{k,l}}{\tau} - \Delta^h \frac{\varphi^{j+1}_{k,l}+\varphi^j_{k,l}}{2} \tag{2.34}$$

et

$$f^j_{k,l} = \frac{1}{h^2}\int_{x_{k-1/2}}^{x_{k+1/2}}\int_{y_{l-1/2}}^{y_{l+1/2}} f(x,\ y,\ t_{j+1/2})\,dx\,dy,$$
$$g^j_{k,l} = \frac{1}{h^2}\int_{x_{k-1/2}}^{x_{k+1/2}}\int_{y_{l-1/2}}^{y_{l+1/2}} g(x,\ y)\,dx\,dy. \tag{2.35}$$

On se ramène au problème

$$\left(E-\frac{\tau}{2}\Delta^h\right)\varphi^{j+1}_{k,l} = \left(E+\frac{\tau}{2}\Delta^h\right)\varphi^j_{k,l} + \tau f^j_{k,l} \text{ dans } D_h\times D_\tau, \tag{2.36}$$

$$\varphi^j_{k,l} = 0 \quad \text{sur} \quad \partial D_h \times D_\tau,$$
$$\varphi^0_{k,l} = g_{k,l} \quad \text{dans} \quad D_h\times\{0\}. \tag{2.37}$$

L'équation (2.36) est résolue formellement par rapport à l'inconnue $\varphi^{j+1}_{k,l}$

$$\varphi^{j+1}_{k,l} = T\varphi^j_{k,l} + \tau S f^j_{k,l}, \tag{2.38}$$

où

$$T = \left(E-\frac{\tau}{2}\Delta^h\right)^{-1}\left(E+\frac{\tau}{2}\Delta^h\right), \qquad S = \left(E-\frac{\tau}{2}\Delta^h\right)^{-1}.$$

La norme de l'opérateur de transition est égale à

$$\|T\| = \max\left\{\left|\frac{1-\frac{4\tau}{h^2}\cos^2\frac{\pi h}{2}}{1+\frac{4\tau}{h^2}\cos^2\frac{\pi h}{2}}\right|,\ \left|\frac{1-\frac{4\tau}{h^2}\sin^2\frac{\pi h}{2}}{1+\frac{4\tau}{h^2}\sin^2\frac{\pi h}{2}}\right|\right\}. \tag{2.39}$$

D'où $\|T\| < 1$.

1.3. Stabilité numérique

Il n'est pas dans nos intentions de donner son entière généralité à la notion de stabilité numérique des schémas aux différences car nous nous intéressons principalement à l'outillage algorithmique simple de l'analyse de leur qualité. De nombreux auteurs ont examiné divers aspects de la théorie de la stabilité et abouti à des généralisations de valeur (Rjabenki, Filippow [VI], Lax [VI], Godounov, Riabenki [VII; 2], Yanenko [III], Isaakson, Keller [III], Richtmyer, Morton [III], Samarski [III; 12]).

Nous illustrerons les définitions et les notions de cette théorie sur l'exemple de plusieurs schémas aux différences. Commençons par le schéma explicite (2.12)

$$\varphi^{j+1} = (E - \tau\Lambda)\varphi^j + \tau f^j, \quad \varphi^0 = g. \tag{3.1}$$

On cherche la solution φ^j pour $0 \leqslant \tau j \leqslant T$.

Supposons que l'opérateur $\Lambda > 0$ possède un système complet de fonctions propres $\{u_n\}$ et l'ensemble de valeurs propres associées $\{\lambda_n > 0\}$ correspondant au problème spectral

$$\Lambda u = \lambda u.$$

Introduisons les séries de Fourier

$$\varphi^j = \sum_n \varphi_n^j u_n, \quad f^j = \sum_n f_n^j u_n, \quad g = \sum_n g_n u_n, \tag{3.2}$$

avec $\varphi_n^j = (\varphi^j, u_n^*)$, $f_n^j = (f^j, u_n^*)$, $g^n = (g, u_n^*)$ (u_n^* étant les fonctions propres du problème spectral adjoint). Portons (3.2) dans (3.1) et multiplions scalairement par u_n^*, il vient, pour les coefficients de Fourier,

$$\varphi_n^{j+1} = (1 - \tau\lambda_n)\varphi_n^j + \tau f_n^j. \tag{3.3}$$

Comme

$$\varphi^0 = \sum_n g_n u_n,$$

on a la condition initiale

$$\varphi_n^0 = g_n. \tag{3.4}$$

On résout (3.3), (3.4) par élimination successive des inconnues. Finalement

$$\varphi_n^j = r_n^j g_n + \tau \sum_{i=1}^{j} r_n^{j-i} f_n^{i-1}, \tag{3.5}$$

où

$$r_n = 1 - \tau\lambda_n. \tag{3.6}$$

L'avant-dernière égalité entraîne pour $\tau > 0$

$$|\varphi_n^j| \leqslant |r_n|^j |g_n| + \tau \sum_{i=1}^{j} |r_n|^{j-i} |f_n^{i-1}|.$$

On renforce cette inégalité en remplaçant $|f_n^{i-1}|$ par $|f_n| = \max_j |f_n^j|$:

$$|\varphi_n^j| \leqslant |r_n|^j |g_n| + \frac{1 - |r_n|^j}{1 - |r_n|} \tau |f_n|. \tag{3.7}$$

On doit à von Neumann le *critère spectral de stabilité* (Richtmyer, Morton [III]) qui dit que si chaque coefficient φ_n^j de la série de Fourier de (3.2) vérifie la relation

$$|\varphi_n^j| \leqslant C_{1n}|g_n| + C_{2n}|f_n|, \quad n = 1, 2, \ldots, \tag{3.8}$$

avec C_{1n}, C_{2n} des constantes uniformément bornées pour $0 \leqslant j\tau \leqslant T$, le schéma aux différences (3.1) est *numériquement stable*. Sous quelles conditions suffisantes imposées aux paramètres du schéma (2.12) on a la relation (3.8)? On montre moyennant (3.7) que ce critère de stabilité est juste avec *)

$$|r_n| < 1, \qquad n = 1, 2, \ldots \tag{3.9}$$

Supposons que l'opérateur Λ a son spectre dans l'intervalle

$$0 < \alpha(\Lambda) \leqslant \lambda_n(\Lambda) \leqslant \beta(\Lambda).$$

Conformément à (3.6), la relation (3.9) a lieu si

$$\tau < \frac{2}{\beta(\Lambda)}. \tag{3.10}$$

Cette inégalité constitue justement une condition constructive de stabilité du schéma aux différences (3.1). Il s'agit d'une condition suffisante et le schéma conserve la propriété voulue pour

$$\tau = \frac{2}{\beta(\Lambda)},$$

auquel cas (3.7) devient

$$|\varphi_n^j| \leqslant |g_n| + j\tau\,|f_n|. \tag{3.11}$$

Mais $j\tau \leqslant T$, T étant fixe, i.e. on prend pour τ faible un nombre important de pas j et $j \to \infty$ quand $\tau \to 0$, mais de façon que la borne supérieure de l'intervalle de temps T reste fixe. Les schémas obtenus sont encore stables au sens de von Neumann.

Passons à des schémas implicites. Dans le cas de (2.13), on a la relation

$$|\varphi_n^j| \leqslant |r_n|^j |g_n| + \frac{1-|r_n|^j}{1-|r_n|}\tau |r_n|\,|f_n| \tag{3.12}$$

analogue à (3.7). Ici

$$r_n = \frac{1}{1+\tau\lambda_n(\Lambda)}.$$

Lorsque $\lambda_n(\Lambda) > 0$, ce schéma aux différences est évidemment stable quel que soit $\tau > 0$ parce que

$$|r_n| < 1, \qquad n = 1, 2, \ldots,$$

et cette stabilité est dite *absolue*.

---

*) Dans la suite nous introduirons une contrainte plus lâche sur la norme de l'opérateur de transition.

S'agissant du schéma de Crank-Nicholson (2.15), les coefficients de Fourier de la solution s'évaluent par

$$|\varphi_n^j| \leqslant |r_n|^j |g_n| + \frac{1-|r_n|^j}{1-|r_n|}\tau|\mu_n||f_n|, \tag{3.13}$$

où

$$r_n = \frac{1-\frac{\tau}{2}\lambda_n(\Lambda)}{1+\frac{\tau}{2}\lambda_n(\Lambda)}, \qquad \mu_n = \frac{1}{1+\frac{\tau}{2}\lambda_n(\Lambda)}.$$

D'où $|r_n| < 1$ avec tout $\tau > 0$ si $\lambda_n(\Lambda) > 0$.

Il y a lieu de noter que, primo, la stabilité au sens de von Neumann relève du spectre de l'opérateur du problème (i.e. le calcul de la plus grande valeur propre ou sa majoration constitue une étape nécessaire de l'algorithme) et, secundo, le critère spectral en question établit la propriété de stabilité par rapport à chaque composante de Fourier en laissant de côté la stabilité pour la norme énergie. Or, il arrive que la norme de φ^j caractérise seule la solution du problème. On a donc essayé d'autres définitions de la stabilité qui soient afférantes aux normes des opérateurs. Il n'est pas moins vrai que le critère local de von Neumann joue un rôle absolument fondamental dans les applications.

Passons à une définition plus générale de la stabilité numérique. Considérons le problème

$$\begin{aligned} &\frac{\partial\varphi}{\partial t} + A\varphi = f \quad \text{dans} \quad D \times D_t, \\ &\varphi = g \quad \text{pour} \quad t = 0 \end{aligned} \tag{3.14}$$

qui est approché par le problème aux différences

$$\varphi^{j+1} = T\varphi^j + \tau S f^j \quad \text{sur} \quad D_h \times D_\tau, \tag{3.15}$$
$$\varphi^{(0)} = g.$$

Nous dirons que le schéma aux différences (3.15) est *stable* si l'on a, pour tout paramètre h caractéristique de la discrétisation et $j \leqslant T/\tau$, la relation

$$\|\varphi^j\|_{\Phi_h} \leqslant C_1 \|g\|_{G_h} + C_2 \|f\|_{F_h}, \tag{3.16}$$

les constantes C_1, C_2 étant uniformément bornées sur $0 \leqslant t \leqslant T$ et indépendantes de τ, h, g et f.

La notion de stabilité numérique est intimement liée à celle des problèmes à argument continu bien posés (Godounov [II], M. M. Lavrentiev [II], Yanenko [III]). On peut dire que la stabilité numérique établit, pour les problèmes à argument discret, la dépendance continue de la solution par rapport aux données initiales.

En effet, soit $f = f_*$, $g = g_*$ les entrées du problème (3.15). Cherchons une solution et notons-la φ_*. Prenons $f = f_* + \xi$, $g = g_* + \delta$ pour nouvelles données initiales et soit φ_{**} la solution correspondante. On a pour la différence $\varepsilon = \varphi_{**} - \varphi_*$

$$\varepsilon^{j+1} = T\varepsilon^j + \tau S\xi^j, \qquad \varepsilon^0 = \delta,$$

et la condition de stabilité s'écrit

$$\| \varepsilon^{j+1} \|_{\Phi_h} \leqslant C_1 \| \delta \|_{G_h} + C_2 \| \xi \|_{F_h}.$$

Il en résulte que les petites variations de f et g entraînent les petites variations de φ.

La définition (3.16) de la stabilité relie, on le constate facilement, la solution même avec l'information à priori sur les données initiales. Dans de nombreux problèmes, elle s'avère plus commode que celle de von Neumann encore que son contenu informationnel soit moins riche. Abordons le schéma (2.12) avec cette optique. On récrit la relation de récurrence (3.1)

$$\varphi^{j+1} = T\varphi^j + \tau f^j, \qquad \varphi^0 = g, \tag{3.17}$$

avec

$$T = E - \tau\Lambda, \tag{3.18}$$

Λ étant l'opérateur approchant A. La solution formelle de (3.17) est de la forme

$$\varphi^{j+1} = T^j g + \tau \sum_{i=1}^{j} T^{j-i} f^{i-1}. \tag{3.19}$$

Evaluons-la en norme, il vient

$$\| \varphi^j \| \leqslant \| T \|^j \| g \| + \tau \sum_{i=1}^{j} \| T \|^{j-i} \| f^{i-1} \|. \tag{3.20}$$

Substituons à $\| f^{i-1} \|$ la valeur maximale par rapport à tous les j de l'intervalle de temps fixe. Soit

$$\| f \| = \max \| f^j \|,$$

auquel cas

$$\| \varphi^j \| \leqslant \| T \|^j \| g \| + \frac{1 - \| T \|^j}{1 - \| T \|} \tau \| f \|. \tag{3.21}$$

Avec l'hypothèse

$$\| T \| < 1, \tag{3.22}$$

le schéma (2.12) est stable au sens de la définition (3.16). La condition (3.22) est naturellement une condition suffisante. On pourrait établir des critères plus fins à l'aide des normes $\| T^i \|$, $i = 1, 2, \ldots$.

. . ., j. Cette stabilité fort générale a été étudiée par Lax et Richtmyer [VII]. Il y a lieu de dire que le fait d'affaiblir la condition nuit à l'aspect constructif du critère de stabilité, si bien qu'on utilise dans la pratique la condition suffisante (3.22).

Plaçons-nous dans le cas $\Lambda = \Lambda^* > 0$ et posons

$$J[\varphi] = \frac{(T\varphi, T\varphi)}{(\varphi, \varphi)} \tag{3.23}$$

(voir (3.18)). Alors

$$J[\varphi] = 1 - 2\tau \frac{(\Lambda\varphi, \varphi)}{(\varphi, \varphi)} + \tau^2 \frac{(\Lambda\varphi, \Lambda\varphi)}{(\varphi, \varphi)}.$$

Soit

$$\varphi = \sum_n \varphi_n u_n,$$

$\{u_n\}$ étant une base de Λ. Dans ce cas

$$J[\varphi] = 1 - 2\tau\widetilde{\lambda} + \tau^2\widetilde{\lambda}^2, \tag{3.24}$$

où

$$\widetilde{\lambda} = \frac{\sum\limits_n \lambda_n(\Lambda)\varphi_n^2}{\sum\limits_n \varphi_n^2}, \qquad \widetilde{\lambda}^2 = \frac{\sum\limits_n [\lambda_n(\Lambda)]^2 \varphi_n^2}{\sum\limits_n \varphi_n^2}.$$

A quelle condition doit être assujettie τ pour que $J[\varphi] \leqslant 1$, i.e.

$$1 - 2\tau\widetilde{\lambda} + \tau^2\widetilde{\lambda}^2 \leqslant 1 ?$$

Cette condition est

$$\tau \leqslant 2\frac{\widetilde{\lambda}}{\widetilde{\lambda}^2} = 2\frac{\sum\limits_n \lambda_n \varphi_n^2}{\sum\limits_n \lambda_n^2 \varphi_n^2}.$$

Si $\beta(\Lambda) = \| \Lambda \| = \max\limits_n \lambda_n(\Lambda) = \lambda_1(\Lambda)$, on en tire

$$\tau \leqslant \frac{2}{\lambda_1(\Lambda)} \frac{\varphi_1^2 + \sum\limits_{n\neq 1} \frac{\lambda_n(\Lambda)}{\lambda_1(\Lambda)} \varphi_n^2}{\varphi_1^2 + \sum\limits_{n\neq 1} \frac{[\lambda_n(\Lambda)]^2}{[\lambda_1(\Lambda)]^2} \varphi_n^2}. \tag{3.25}$$

Puisque

$$\frac{\varphi_1^2 + \sum\limits_{n\neq 1} \frac{\lambda_n(\Lambda)}{\lambda_1(\Lambda)} \varphi_n^2}{\varphi_1^2 + \sum\limits_{n\neq 1} \frac{[\lambda_n(\Lambda)]^2}{[\lambda_1(\Lambda)]^2} \varphi_n^2} \geqslant 1,$$

on aboutit à la condition suffisante de $J[\varphi] \leqslant 1$:

$$\tau \leqslant \frac{2}{\beta(\Lambda)}.$$

Par définition (1.11) de la norme d'un opérateur, on a

$$\| T \|^2 = \sup_{\varphi} (J[\varphi]) \leqslant 1,$$

et il y a donc stabilité numérique au sens de (3.16). S'agissant d'un opérateur auto-adjoint, il y a coïncidence de la condition suffisante de stabilité numérique d'après von Neumann et de celle au sens de la définition (3.16). Le lien entre les deux définitions est analysé dans les monographies de Godounov et Riabenki [III; 33, 34] et de Richtmyer et Morton [III].

On traite de façon analogue la stabilité des équations implicites (2.13) et (2.15) pour lesquelles on a

$$\| \varphi^j \| \leqslant \| T \|^j \| g \| + \frac{1 - \| T \|^j}{1 - \| T \|} \tau \| S \| \| f \|,$$

$T = (E + \tau\Lambda)^{-1}$, $S = (E + \tau\Lambda)^{-1}$ pour (2.13) ; $T = \left(E + \frac{\tau}{2}\Lambda\right)^{-1} \left(E - \frac{\tau}{2}\Lambda\right)$, $S = \left(E + \frac{\tau}{2}\Lambda\right)^{-1}$ dans l'autre cas.

On montre aisément leur stabilité absolue au sens de (3.16) si $\Lambda > 0$.

Quelques mots sur les passages à la limite. Certains analogues discrétisés de problèmes d'évolution de la physique mathématique le sont en les variables d'espace et de temps à la fois, i.e. l'opérateur de transition $T = T(\tau, h)$ dépend des deux paramètres.

Dans le cas de la discrétisation complète, on construit en général un algorithme stable par le seul fait d'établir une relation entre τ et h qui garantisse la stabilité numérique. Si le schéma aux différences s'avère stable pour tous les $\tau > 0$ et $h > 0$, il l'est *absolument*, et si la stabilité n'a lieu que pour une relation déterminée entre les pas, le schéma est dit *conditionnellement stable.*

Supposons qu'on se donne une dépendance de la forme

$$\tau \leqslant Ch^p, \tag{3.26}$$

avec C et p des constantes données indépendantes de τ et h. On observe que ces relations interviennent d'ordinaire lorsqu'on analyse l'amplitude de la plus « courte » perturbation et qu'elles expriment de règle le lien entre la plus faible étendue et la plus courte durée de phénomènes que nous voulons décrire par le schéma aux différences.

Les perturbations plus importantes (de l'ordre de plusieurs h) sont évidemment décrites plus exactement.

Supposons que nous voulons améliorer la précision de la solution en diminuant formellement le pas h. On doit en même temps diminuer τ afin que l'inégalité indiquée reste en vigueur pour ces paramètres nouveaux. Cela signifie qu'on peut passer à la limite pour $\tau \to 0$ et $h \to 0$ en récrivant (3.26) par exemple de façon suivante

$$\frac{\tau}{h^{\nu}} = \text{const} \leqslant C.$$

On trouve dans la littérature mathématique d'autres définitions de la stabilité numérique. On dit par exemple que le schéma est *stable* si

$$\| T \| \leqslant 1 + O(\tau). \tag{3.27}$$

Avec cette définition, les erreurs d'arrondi peuvent croître exponentiellement dans le temps pour τ petits (Yanenko [III], Rojdestvenski et Yanenko [II, 7]). Les notions de stabilité de Strang [VI], Godounov et Riabenki [III ; 33, 34], Kreiss [VI], Yanenko et Chokin [VII ; 8], Samarski [III ; 12] permettent d'étendre la classe de schémas aux différences utiles dans les applications.

Nous avons étudié la stabilité numérique en supposant l'opérateur Λ indépendant du temps. Cette hypothèse est naturelle pour plusieurs problèmes de la physique mathématique et elle permet de plus d'introduire des procédés constructifs très employés en calcul numérique. En effet, l'étude de la stabilité se ramène à évaluer la norme de T. Il est connu dès le numéro 1.1 que le carré de la norme de T coïncide avec le rayon spectral de l'opérateur positif auto-adjoint T^*T et on définit le rayon spectral par le processus itératif

$$\| T \|^2 = \lim_{k \to \infty} \frac{(T^*T\varphi^{(k)}, \varphi^{(k)})}{(\varphi^{(k)}, \varphi^{(k)})},$$

$\varphi^{(k)}$ étant éléments de

$$\varphi^{(k+1)} = \frac{1}{\| \varphi^{(k)} \|} T^*T\varphi^{(k)}. \tag{3.28}$$

Ainsi, le problème de définir la norme de T se réduit à une suite de relations de récurrence (3.28). C'est le procédé le mieux élaboré en vue du calcul sur machine. Lorsque T est auto-adjoint, on a

$$\| T \| = \beta(T).$$

Faisons plusieurs remarques d'ordre particulier. L'étude de la stabilité de schémas aux différences se fait des fois par définition du rayon spectral d'un problème infini périodique par rapport aux variables d'espace. Si les conditions aux limites sont non périodiques, on évalue nécessairement le rayon spectral par la méthode de Kellogg pour T construits compte tenu des conditions aux limites réelles.

Si Λ varie avec le temps, le problème de stabilité se complique singulièrement car il en est également de la norme de T et, en général, on cherche le rayon spectral à chaque pas vu qu'il dépend lui aussi du numéro du pas temporel. Il y a intérêt dans ce cas à s'adresser à des problèmes aux différences absolument stables (voir Chapitre 4).

On note pour conclure que si l'équation d'évolution est approchée dans les espaces des fonctions discrètes définies sur $D_h \times D_\tau$, il serait bon de définir la stabilité en termes de mêmes espaces. En effet, soit le problème aux différences

$$\begin{aligned} L^{h\tau}\varphi^{h\tau} &= f^{h\tau} \quad \text{dans} \quad D_h \times D_\tau, \\ l^{h\tau}\varphi^{h\tau} &= g^{h\tau} \quad \text{sur} \quad \partial D_h \times D_\tau. \end{aligned} \tag{3.29}$$

Prenons pour critère de stabilité

$$\| \varphi^{h\tau} \|_{\Phi_{h\tau}} \leqslant C_1 \| f^{h\tau} \|_{F_{h\tau}} + C_2 \| g^{h\tau} \|_{G_{h\tau}}, \tag{3.30}$$

avec C_1, C_2 des constantes indépendantes de h, τ, $f^{h\tau}$, $g^{h\tau}$ sur $0 \leqslant t \leqslant T$.

Cette approche générale de la stabilité a été proposée par Rjabenki, Filippow [VI].

Supposons que le problème de la physique mathématique est approché par une équation aux différences de façon que celle-ci tienne compte des conditions aux limites. On introduit le critère de stabilité commode

$$\| \varphi^{h\tau} \|_{\Phi_{h\tau}} \leqslant C \| f^{h\tau} \|_{F_{h\tau}}, \tag{3.31}$$

C étant bornée sur $0 \leqslant t \leqslant T$.

1.4. Un théorème de convergence

Filippov définit la stabilité de problèmes aux différences arbitraires

$$L^{h\tau}\varphi^{h\tau} = f^{h\tau}$$

comme la borne uniforme de l'opérateur $(L^{h\tau})^{-1}$. Il démontre que l'approximation et la stabilité entraînent la convergence de la solution du problème aux différences vers celle du problème différentiel (Filippov [VI ; 6]). Lax a proposé pour les problèmes d'évolution bien posés les définitions de l'approximation et de la stabilité telles que s'il y a approximation, la stabilité implique la convergence (voir Richtmyer, Morton [III]). Ce résultat porte le nom de théorème de Lax.

Qu'il s'agisse de problèmes stationnaires ou de cas d'évolution, on s'inspire des mêmes idées pour établir la convergence. Prenons donc le problème stationnaire

$$\begin{aligned} A\varphi &= f \quad \text{dans} \quad D, \\ a\varphi &= g \quad \text{sur} \quad \partial D \end{aligned} \tag{4.1}$$

qui est approché par le schéma aux différences

$$\begin{aligned} A^h\varphi^h &= f^h \quad \text{dans} \quad D_h, \\ a^h\varphi^h &= g^h \quad \text{sur} \quad \partial D_\tau. \end{aligned} \tag{4.2}$$

On a le *théorème de convergence* suivant: *on suppose que*

1) *le schéma aux différences* (4.2) *approche le problème initial* (4.1) *à l'ordre n par rapport à la solution* φ;

2) A^h *et* a^h *sont des opérateurs linéaires*;

3) *il y a stabilité de* (4.2) *au sens de* (3.30), *i.e. on trouve des constantes positives* $\overline{h}$, C_1, C_2 *telles que le problème* (4.2) *admette pour tous les* $h < \overline{h}$, $f^h \in F_h$, $g^h \in G_h$ *une solution* φ^h *(unique) qui vérifie l'inégalité*

$$\| \varphi^h \|_{\Phi_h} \leqslant C_1 \| f^h \|_{F_h} + C_2 \| g^h \|_{G_h}. \tag{4.3}$$

La solution φ^h *du problème aux différences converge vers la solution* φ *du problème initial, i.e.* $\lim\limits_{h \to 0} \| (\varphi)_h - \varphi^h \|_{\Phi_h} = 0$, *et la vitesse de convergence s'évalue par*

$$\| (\varphi)_h - \varphi^h \|_{\Phi_h} \leqslant (C_1 M_1 + C_2 M_2)\, h^n, \tag{4.4}$$

M_1 *et* M_2 *étant les constantes de* (2.3).

DÉMONSTRATION. Soit $\overline{h}$ la plus petite des $\overline{h}$ des définitions de l'approximation et de la stabilité. Etant donnée la stabilité, il existe pour n'importe quels seconds membres f^h et g^h, $h < \overline{h}$, une solution unique φ^h, i.e. on peut considérer la différence $(\varphi)_h - \varphi^h$, $h < \overline{h}$.

A^h étant linéaire, on a

$$A^h [(\varphi)_h - \varphi^h] = A^h (\varphi)_h - A_h \varphi^h = A_h (\varphi)_h - f^h.$$

De même,

$$a^h [(\varphi)_h - \varphi^h] = a^h (\varphi)_h - g^h.$$

Par suite de la stabilité et de l'approximation,

$$\begin{aligned} A^h [(\varphi)_h - \varphi^h] &= A^h (\varphi)_h - f^h, \\ a^h [(\varphi)_h - \varphi^h] &= a^h (\varphi)_h - g^h, \end{aligned} \tag{4.5}$$

entraînent, vu que $h < \overline{h}$,

$$\| (\varphi)_h - \varphi^h \|_{\Phi_h} \leqslant C_1 \| A^h (\varphi)_h - f^h \|_{F_h} + C_2 \| a^h (\varphi)_h - g^h \|_{G_h} \leqslant$$
$$\leqslant C_1 M_1 h^{n_1} + C_2 M_2 h^{n_2} \leqslant (C_1 M_1 + C_2 M_2)\, h^n.$$

On estime, sans restreindre la généralité, que $h < 1$, ce qui achève la démonstration.

On rappelle que la propriété de linéarité n'a été supposée juste que pour A^h et a^h. Godunov et Rjabenki [III] ont étendu la définition de la stabilité au cas non linéaire et donné un théorème de convergence correspondant.

Plaçons-nous maintenant dans un problème d'évolution. On a

$$\left.\begin{aligned} \delta f^{h\tau} &\equiv L^{h\tau}[(\varphi)_{h\tau} - \varphi^{h\tau}] \equiv L^{h\tau}(\varphi)_{h\tau} - f^{h\tau}, \\ \delta g^{h\tau} &\equiv l^{h\tau}[(\varphi)_h - \varphi^{h\tau}] \equiv l^{h\tau}(\varphi)_{h\tau} - g^{h\tau}. \end{aligned}\right\} \tag{4.6}$$

On trouve à partir de la condition de stabilité (3.30)

$$\|(\varphi)_{h\tau} - \varphi^{h\tau}\|_{\Phi_h} \leqslant C_1 \|\delta f^{h\tau}\|_{F_{h\tau}} + C_2 \|\delta g^{h\tau}\|_{G_{h\tau}},$$

ou, compte tenu de (2.17),

$$\|(\varphi)_{h\tau} - \varphi^{h\tau}\|_{\Phi_{h\tau}} \leqslant K_1 h^n + K_2 \tau^p, \tag{4.7}$$

où

$$K_1 = C_1 M_1 + C_2 M_2, \qquad K_2 = C_1 N_1 + C_2 N_2.$$

L'estimation (4.7) prouve que la solution aux différences converge vers la solution exacte et elle donne une idée nette de la convergence par rapport au pas d'espace h et au pas de temps τ.

Le théorème ci-dessus assujettit C_1 et C_2 à la condition sévère d'être indépendantes de h et τ. C'est la première contrainte qui est particulièrement désagréable. Il arrive cependant que ces constantes tendent vers l'infini avec $h \to 0$. Soit

$$C_1^h = \frac{C_1}{h^m} \quad \text{et} \quad C_2^h = \frac{C_2}{h^m},$$

$m \geqslant 0$. On évalue alors la convergence par

$$\|\varphi^{h\tau} - \varphi_{h\tau}\|_{\Phi_{h\tau}} \leqslant M h^{k-m} + N \tau^p h^{-m}.$$

Si $k > m$ et $\tau^p h^{-m} \to 0$ pour $\tau \to 0$, $h \to 0$, il y a convergence. On énonce naturellement un théorème de convergence si C_1 et C_2 dépendent de h et τ (Strang [VI] [VII]).

CHAPITRE 2

MÉTHODES DE CONSTRUCTION DE SCHÉMAS AUX DIFFÉRENCES POUR LES ÉQUATIONS DIFFÉRENTIELLES

La construction d'équations aux différences pour les problèmes de la physique mathématique repose sur divers principes. La question est particulièrement bien étudiée si la solution et les coefficients des équations sont suffisamment réguliers, et on trouve alors des schémas d'ordre d'approximation élevé. Ces schémas deviennent un instrument de travail préféré car la science et la technique se posent des problèmes complexes et laborieux avant que le matériel de calcul approprié fasse son apparition. Il paraît donc utile d'approcher des fois tel problème avec une précision donnée non pas en augmentant formellement la dimension des sous-espaces (en réduisant par exemple le pas du réseau), mais à partir de l'information à priori sur la régularité de la solution *). Cette façon de procéder est très prometteuse dans plusieurs cas. Elle a permis de construire des équations aux différences par des techniques commodes et assez universelles qui s'inspirent des méthodes variationnelles de Ritz, de Galerkin et des moindres carrés. On note néanmoins que la classe de problèmes à solutions régulières est plutôt restreinte, si bien qu'on doit s'intéresser en premier lieu aux méthodes pour les problèmes à coefficients discontinus qui interviennent par exemple en théorie de la diffusion, dans la propagation de la chaleur, en hydrodynamique.

Aussi nous passerons sur certains résultats originaux de caractère fort général relatifs à la formation d'équations aux différences de haute précision pour donner en revanche une idée générale de la discrétisation d'équations dont les solutions ne sont pas nécessairement très régulières. Tous les procédés employés se généralisent naturellement aux problèmes à données et solutions régulières.

Pour mettre en relief diverses voies de recherche dans le domaine considéré nous nous adressons d'abord aux problèmes aux limites pour les équations différentielles ordinaires pour exposer ensuite des approches plus ou moins générales des problèmes bi et n-dimensionnels. Les références à la Bibliographie *in fine* permettront (nous

*) Voir également Chapitre 6.

l'espérons au moins) aux lecteurs d'approfondir leurs connaissances des aspects théoriques et algorithmiques de la question.

2.1. Méthode de l'identité intégrale

Il existe nombre de méthodes de résolution des problèmes aux limites pour les équations différentielles ordinaires. Nous allons en examiner quelques-unes parmi les plus connues. Notre but étant de familiariser le lecteur avec certains principes de l'analyse numérique, nous nous bornerons à des exemples simples qui illustrent au mieux les procédés étudiés.

Les méthodes de construction de problèmes aux limites discrets pour les équations différentielles ordinaires à coefficients réguliers constituent un chapitre bien élaboré des mathématiques. Quant à nous, c'est le cas des coefficients discontinus qui nous intéresse. Les problèmes de ce type interviennent dans beaucoup d'applications importantes. Les premiers schémas ont été examinés par Tikhonov et Samarski [IV; 27].

Le présent paragraphe est consacré à la construction d'équations discrétisées de la diffusion à l'aide d'une identité intégrale obtenue par l'auteur [XVII; 10].

Considérons l'équation de diffusion pour les domaines à une dimension

$$-\frac{d}{dx}\, p\, \frac{d\varphi}{dx} + q\varphi = f, \tag{1.1}$$

avec $p = p(x) \geqslant p_0 > 0$ le coefficient de diffusion, $q = q(x) \geqslant 0$ la section d'absorption pour les particules et $f = f(x)$ les sources de substance diffusante. On estime que p, q et f sont des fonctions continues par morceaux sur $[0, 1]$ avec peut-être des points de discontinuité de première espèce.

On demande la solution de (1.1) possédant le « flux » dérivable

$$J = p\, \frac{d\varphi}{dx}$$

avec les conditions aux limites

$$\varphi(0) = 0, \quad \varphi(1) = 0. \tag{1.2}$$

Soit, sur l'intervalle de variation de x, deux systèmes de nœuds : $\{x_k\}$ (système principal) et $\{x_{k+1/2}\}$ (système auxiliaire). Les points de $\{x_k\}$ et de $\{x_{k+1/2}\}$ alternent, i.e. $x_k < x_{k+1/2} < x_{k+1}$. On suppose que $x_{k+1/2} = \frac{x_{k+1} + x_k}{2}$, $0 = x_0 < x_{1/2} < \ldots < x_{n-1/2} < x_n = 1$.

Intégrons (1.1) par rapport à x de $x_{k-1/2}$ à $x_{k+1/2}$, il vient la relation du bilan

$$-J_{k+1/2}+J_{k-1/2}+\int_{x_{k-1/2}}^{x_{k+1/2}}(q\varphi-f)\,dx=0, \tag{1.3}$$

où

$$J_{k\pm1/2}=J(x_{k\pm1/2}).$$

Cherchons $J_{k\pm1/2}$. On intègre (1.1) sur l'intervalle $(x_{k-1/2}, x)$

$$p\frac{d\varphi}{dx}=J_{k-1/2}+\int_{x_{k-1/2}}^{x}(q\varphi-f)\,d\xi, \tag{1.4}$$

on divise par p et on intègre de x_{k-1} à x_k, il vient

$$\varphi_k-\varphi_{k-1}=J_{k-1/2}\int_{x_{k-1}}^{x_k}\frac{dx}{p}+\int_{x_{k-1}}^{x_k}\frac{dx}{p}\int_{x_{k-1/2}}^{x}(q\varphi-f)\,d\xi. \tag{1.5}$$

Résolvons cette équation en $J_{k-1/2}$:

$$J_{k-1/2}=\frac{1}{\int_{x_{k-1}}^{x_k}\frac{dx}{p}}\left[\varphi_k-\varphi_{k-1}-\int_{x_{k-1}}^{x_k}\frac{dx}{p}\int_{x_{k-1/2}}^{x}(q\varphi-f)\,d\xi\right]. \tag{1.6}$$

Par substitution $k=k+1$ dans (1.6) on obtient une expression analogue de $J_{k+1/2}$. On a donc déterminé les flux $J_{k\pm1/2}$ par les fonctions connues et la solution. La relation (1.6) est exacte. Portons (1.6) et l'expression de $J_{k+1/2}$ dans (1.3), il vient

$$-\frac{\varphi_{k+1}-\varphi_k}{\int_{x_k}^{x_{k+1}}\frac{dx}{p}}+\frac{\varphi_k-\varphi_{k-1}}{\int_{x_{k-1}}^{x_k}\frac{dx}{p}}+\int_{x_{k-1/2}}^{x_{k+1/2}}(q\varphi-f)\,dx=$$

$$=-\frac{1}{\int_{x_k}^{x_{k+1}}\frac{dx}{p}}\int_{x_k}^{x_{k+1}}\frac{dx}{p}\int_{x_{k+1/2}}^{x}(q\varphi-f)\,d\xi+$$

$$+\frac{1}{\int_{x_{k-1}}^{x_k}\frac{dx}{p}}\int_{x_{k-1}}^{x_k}\frac{dx}{p}\int_{x_{k-1/2}}^{x}(q\varphi-f)\,d\xi, \tag{1.7}$$

qui constitue l'*identité fondamentale* permettant d'obtenir des équations aux différences finies.

Introduisons l'opérateur A défini sur la classe Φ de solutions de (1.1) par l'égalité

$$(A\varphi)_k = -\frac{1}{\Delta x_k}\left(\frac{\varphi_{k+1}-\varphi_k}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}} - \frac{\varphi_k-\varphi_{k-1}}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}} - \int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\varphi\,dx - \right.$$

$$\left. - \frac{1}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p} \int\limits_{x_{k+1/2}}^{x} q\varphi\,d\xi + \frac{1}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} \int\limits_{x_{k-1/2}}^{x} q\varphi\,d\xi \right),$$

et le vecteur f de composantes *)

$$(f)_k = \frac{1}{\Delta x_k}\int\limits_{x_{k-1/2}}^{x_{k+1/2}} f\,dx +$$

$$+ \frac{1}{\Delta x_k}\left(\frac{1}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p} \int\limits_{x_{k+1/2}}^{x} f\,d\xi - \frac{1}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} \int\limits_{x_{k-1/2}}^{x} f\,d\xi \right), \quad (1.8)$$

avec

$$\Delta x_k = x_{k+1/2} - x_{k-1/2}, \qquad k = 1, \ldots, n-1.$$

Nous supposerons, pour simplifier l'exposé, que les solutions de (1.1) sont de classe Φ formée de fonctions ayant une certaine régularité et vérifiant les conditions aux limites $\varphi = 0$. Groupons les relations (1.7) pour $k = 1, \ldots, n-1$, il vient le système

$$A\varphi = f. \quad (1.9)$$

Considérons diverses approximations de (1.9). Munissons l'espace F_h de fonctions discrètes de la forme $\varphi^h = (\varphi_1^h, \ldots, \varphi_{n-1}^h)$ définies aux points $x_1, x_2, \ldots, x_{n-1}$ de la norme

$$\|\varphi^h\|_{F_h}^2 = \sum_{k=1}^{n-1} \varphi_k^2 \Delta x_k. \quad (1.10)$$

*) Ne pas confondre avec $f(x_k)$!

Soit le problème approché

$$A^h\varphi^h = f^h, \tag{1.11}$$

où

$$(A^h\varphi^h)_k = -\frac{1}{\Delta x_k}\left(\frac{\varphi^h_{k+1}-\varphi^h_k}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}} - \frac{\varphi^h_k-\varphi^h_{k-1}}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}} - \varphi^h_k\int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\,dx\right),$$
$$(f^h)_k = \frac{1}{\Delta x_k}\int\limits_{x_{k-1/2}}^{x_{k+1/2}} f\,dx \tag{1.12}$$

pour $k = 1, \ldots, n-1$, et $\varphi^h_0 = \varphi^h_n = 0$. Utilisons l'inégalité triangulaire, on a par définition de l'approximation *)

$$\|A^h(\varphi)_h - f^h\|_{F_h} \leqslant \|\xi^h\|_{F_h} + \|\eta^h\|_{F_h} + \|\Theta^h\|_{F_h}, \tag{1.13}$$

où

$$(\xi^h)_k = \frac{1}{\Delta x_k}\left(\int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\varphi\,dx - \varphi_k\int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\,dx\right),$$

$$(\eta^h)_k =$$

$$= -\frac{1}{\Delta x_k}\left(\frac{1}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}}\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}\int\limits_{x_{k+1/2}}^{x} q\varphi\,d\xi - \frac{1}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}}\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}\int\limits_{x_{k-1/2}}^{x} q\varphi\,d\xi\right),$$

$$(\Theta^h)_k =$$

$$= -\frac{1}{\Delta x_k}\left(\frac{1}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}}\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}\int\limits_{x_{k+1/2}}^{x} f\,d\xi - \frac{1}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}}\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}\int\limits_{x_{k-1/2}}^{x} f\,d\xi\right).$$

Evaluons $\|\xi^h\|_{F_h}$, $\|\eta^h\|_{F_h}$ et $\|\Theta^h\|_{F_h}$. Faisons l'hypothèse: q, $f \in Q^{(2)}(0, 1)$, $p \in Q^{(3)}(0, 1)$, $Q^{(s)}(0, 1)$ étant l'espace des fonctions continues par morceaux sur (0,1), ainsi que leurs dérivées jusqu'à l'ordre s inclus, qui présentent peut-être des discontinuités de première espèce en $0 < y_1 < y_2 < \ldots < y_m < 1$. L'ensemble des points $\{y_l\}_{l=1}^m$ sera supposé appartenir à l'ensemble des nœuds

*) $(u)_h$ désignera partout, pour toute u continue sur [0, 1], un vecteur $(n-1)$-dimensionnel $\in F_h$ de composantes $u(x_k)$.

$\{x_k\}_{k=1}^{n-1}$, condition nécessaire dans l'analyse des erreurs d'approximation.

Les hypothèses faites entraînent la continuité de la solution φ du problème (1.1), φ étant continue sur chaque $[y_l, y_{l+1}]_{l=1}^{m-1}$ ainsi que toutes ses dérivées d'ordre 4 inclus, i.e. $\varphi \in Q^{(4)}(0, 1)$. Voyons le comportement des composantes des ξ^h, η^h et Θ^h si $h \ll 1$, $h = \max\limits_{0 \leqslant k \leqslant n-1} |x_{k+1} - x_k|$. En développant en série de Taylor usuelle au voisinage des nœuds, on montre facilement que les modules des composantes sont majorés par les composantes correspondantes de ω^h, où

$$(\omega^h)_k = \begin{cases} Nh \text{ si } x_k \text{ se confond avec l'un des } y_l,\ l = 1, \ldots, m, \\ M(|\Delta x_{k+1/2} - \Delta x_{k-1/2}| + h^2) \text{ dans les cas restants,} \end{cases}$$

M et N étant des constantes positives. Nous avons introduit la notation $\Delta x_{k+1/2} = x_{k+1} - x_k$. En effet, admettons que le domaine de définition de la solution renferme un point de discontinuité des coefficients. Soit $x = x_l$ ce point. On écrit par suite de (1.10)

$$\| \omega^h \|_{F_h}^2 = \sum_{\substack{k=1 \\ k \neq l}}^{n-1} (\omega^h)_k^2 \Delta x_k + (\omega^h)_l^2 \Delta x_l.$$

Supposons ensuite que $\widetilde{h} = \max \{\Delta x_k,\ 2(1 - x_{n-1/2}),\ 2x_{1/2}\}$. Etant donnée la relation

$$1 - \widetilde{h} \leqslant \sum_{k=1}^{n-1} \Delta x_k < 1,$$

les estimations locales ci-dessus de $(\omega^h)_k$ entraînent

$$\| \omega^h \|_{F_h}^2 \leqslant M^2 h^4 + N^2 h^3.$$

D'où

$$\| \omega^h \|_{F_h} \leqslant C h^{3/2}.$$

Ainsi, les normes des vecteurs erreurs de discrétisation ξ^h, η^h et Θ^h s'évaluent par

$$\max (\| \xi^h \|_{F_h},\ \| \eta^h \|_{F_h},\ \| \Theta^h \|_{F_h}) \leqslant C h^{3/2} \tag{1.14}$$

avec C une constante positive indépendante de h, si, ou bien le réseau est régulier sur chacun des $[0, y_1]$, $[y_1, y_2]$, ..., $[y_m, 1]$, ou bien il est quasi régulier, i.e. $|\Delta x_{k+1/2} - \Delta x_{k-1/2}| \leqslant Ch^2$, C une constante positive, n'a lieu pour $h \to 0$ qu'un nombre fini de fois. On pourrait continuer la liste des conditions, mais celles citées ont une importance pratique particulièrement grande.

Avec p, q ou f d'ordre de régularité inférieur d'une unité, on a l'estimation

$$\max\left(\|\xi^h\|_{F_h},\quad \|\eta^h\|_{F_h},\quad \|\Theta^h\|_{F_h}\right)\leqslant C_1 h.$$

Le schéma (1.11) est d'usage peu fréquent parce que l'intégration explicite de p, q et f s'avère quelquefois difficile. On recourt en général à la variante simplifiée

$$(A^h\varphi^h)_k=-\frac{1}{\Delta x_k}\left\{p_{k+1/2}\frac{\varphi_{k+1}-\varphi_k}{\Delta x_{k+1/2}}-p_{k-1/2}\frac{\varphi_k-\varphi_{k-1}}{\Delta x_{k-1/2}}-(q\Delta x)_k\varphi_k\right\},$$

$$(f^h)_k=\frac{1}{\Delta x_k}(f\Delta x)_k=f_k=\frac{f_{k+1/2}(x_k-x_{k-1/2})+f_{k-1/2}(x_{k+1/2}-x_k)}{x_{k+1/2}-x_{k-1/2}},$$

$$k=1,\ 2,\ \ldots,\ n-1.$$

Dans ce cas simple toutes les déductions relatives à la grandeur de l'erreur de discrétisation restent en vigueur à condition de conserver l'hypothèse correspondante sur la régularité des données du problème.

Passons au problème de convergence de la solution aux différences de (1.11)-(1.12) vers la solution exacte du problème initial (1.1) dans les hypothèses faites sur la régularité de p, q et f. Il suffit d'établir la stabilité du schéma (1.11) et d'utiliser le théorème de convergence.

On évalue par l'inégalité de Cauchy-Bouniakovski

$$(\varphi^h,\ f^h)\leqslant\|\varphi^h\|_{F_h}\|f^h\|_{F_h},\tag{1.15}$$

où le produit scalaire est compris au sens de

$$(\varphi,\ \psi)=\sum_{n=1}^{n-1}\Delta x_k\varphi_k\psi_k\qquad(\varphi,\ \psi\in F_h).$$

Voyons de plus près le premier membre de (1.15). Comme $q\ (x)\geqslant 0$ et $p\ (x)\geqslant p_0>0$ par hypothèse, on a

$$(\varphi^h,\ f^h)=(\varphi^h,\ A^h\varphi^h)=$$

$$=\sum_{k=1}^{n}\frac{(\varphi_k-\varphi_{k-1})^2}{\int_{x_{k-1}}^{x_k}\frac{dx}{p}}+\sum_{k=1}^{n-1}\varphi_k^2\int_{x_{k-1/2}}^{x_{k+1/2}}q\,dx\geqslant p_0\sum_{k=1}^{n}\frac{(\varphi_k-\varphi_{k-1})^2}{\Delta x_{k-1/2}}>0,\tag{1.16}$$

avec φ^h un vecteur non nul parce que solution du problème non homogène (1.11) de matrice non dégénérée A^h.

Etant donné que $\varphi_0=\varphi_n=0$, on écrit

$$\varphi_k=\sum_{j=1}^{k}(\varphi_j-\varphi_{j-1})=\sum_{j=1}^{k}\frac{\varphi_j-\varphi_{j-1}}{\sqrt{\Delta x_{j-1/2}}}\sqrt{\Delta x_{j-1/2}},$$

d'où pour les sommes moyennant l'inégalité de Cauchy-Bouniakovski

$$\varphi_k^2 = \Big(\sum_{j=1}^{k} \frac{\varphi_j - \varphi_{j-1}}{\sqrt{\Delta x_{j-1/2}}} \sqrt{\Delta x_{j-1/2}}\Big)^2 \leqslant$$

$$\leqslant \Big(\sum_{j=1}^{k} \frac{(\varphi_j - \varphi_{j-1})^2}{\Delta x_{j-1/2}}\Big)\Big(\sum_{j=1}^{k} \Delta x_{j-1/2}\Big) \leqslant \sum_{j=1}^{n} \frac{(\varphi_j - \varphi_{j-1})^2}{\Delta x_{j-1/2}}.$$

Enfin,

$$\sum_{k=1}^{n-1} \varphi_k^2 \Delta x_k \leqslant \sum_{j=1}^{n} \frac{(\varphi_j - \varphi_{j-1})^2}{\Delta x_{j-1/2}} \sum_{k=1}^{n-1} \Delta x_k < \sum_{j=1}^{n} \frac{(\varphi_j - \varphi_{j-1})^2}{\Delta x_{j-1/2}}, \qquad (1.17)$$

du moment que $\sum_{k=1}^{n-1} \Delta x_k < 1$.

On tire de (1.15)-(1.17)

$$\| \varphi^h \|_{F_h} \cdot \| f^h \|_{F_h} \geqslant (\varphi^h, f^h) \geqslant p_0 \sum_{j=1}^{n} \frac{(\varphi_j - \varphi_{j-1})^2}{\Delta x_{j-1/2}} \geqslant p_0 \| \varphi^h \|_{F_h}^2,$$

si bien que

$$\| \varphi^h \|_{F_h} \leqslant \frac{1}{p_0} \| f^h \|_{F_h},$$

inégalité qui constitue par définition la stabilité de l'algorithme aux différences. Le théorème de convergence (pour la norme (1.10)) fournit donc la majoration

$$\| \varepsilon^h \|_{F_h} \leqslant K h^{3/2} \quad (\varepsilon^h = (\varphi)_h - \varphi^h),$$

$K \geqslant \frac{3C}{p_0}$ étant une constante positive *).

La technique décrite a beau être importante dans la pratique, nous n'en avons besoin que pour la comparer avec les tactiques de discrétisation nouvelles.

2.2. Méthodes variationnelles en physique mathématique

Considérons la fonctionnelle simple (voir Smirnov [V])

$$J(u) = \int_{x_0}^{x_1} \pi(x, u, u')\, dx, \qquad (2.1)$$

*) S'agissant de notre cas, c'est à vrai dire une estimation grossière. L'analyse plus fine montre que $\| \varepsilon^h \|_{F_h} \leqslant Mh^2$ (Tikhonov et Samarski [IV; 27]).

avec $\pi = \pi(x, y, z)$ une fonction donnée continue dans E_3 ainsi que ses dérivées premières et secondes par rapport à x, y, z (E_3 est l'espace euclidien de dimension 3, y et z dépendent de la variable x).

Supposons la fonction $u = u(x)$ prendre des valeurs données aux extrémités de $[x_0, x_1]$

$$u(x_0) = u_0, \quad u(x_1) = u_1, \tag{2.2}$$

$u(x)$ étant continue sur $[x_0, x_1]$ et y admet une dérivée continue $u'(x)$.

Définissons l'ε-voisinage de $u = u(x)$ comme étant une famille de $u_1(x)$ qui vérifient sur $[x_0, x_1]$ tout entier l'inégalité

$$|u_1(x) - u(x)| \leqslant \varepsilon. \tag{2.3}$$

Etablissons la condition sous laquelle la fonctionnelle $J(u)$ possède un extrémum dans l'ε-voisinage. On prend à cet effet une fonction $\eta(x)$ telle que

$$\eta(x_0) = \eta(x_1) = 0, \tag{2.4}$$

et on construit

$$u_\alpha(x) = u(x) + \alpha\eta(x),$$

avec α un petit paramètre. Portons $u_\alpha(x)$ dans la fonctionnelle J, il vient

$$J(u_\alpha) = \int_{x_0}^{x_1} \pi(x, u(x) + \alpha\eta(x), \quad u'(x) + \alpha\eta'(x))\, dx.$$

Développons la fonction sous $\int$ en série de Taylor suivant α. On se borne à des infiniment petits d'ordre 1 et on a

$$J(u_\alpha) = J(u) + \alpha J'(u). \tag{2.5}$$

La quantité

$$\delta J(u) = \int_{x_0}^{x_1} (\pi_u \eta + \pi_{u'} \eta')\, dx \tag{2.6}$$

s'appelle *variation première* de la fonctionnelle J. La formule (2.5) régit la liaison entre la variation de J, la quantité à varier α et la direction de la variation $\eta(x)$. L'intégration par parties de (2.6) donne

$$\delta J(u) = \int_{x_0}^{x_1} \eta(x) \left[\pi_u - \frac{d}{dx} \pi_{u'}\right] dx = 0 \tag{2.7}$$

(nous utilisons les conditions (2.4)).

Etant donné l'arbitraire de $\eta(x)$, on conclut que la courbe $u(x)$ qui réalise l'extrémum de (2.1) vérifie l'équation différentielle

$$\pi_u - \frac{d}{dx}\pi_{u'} = 0 \tag{2.8}$$

qu'on appelle *équation d'Euler.*

Nous nous sommes placé dans le cas simple d'une fonction u et d'une variable indépendante x. Des cas plus complexes se traitent de façon analogue. Soit par exemple la fonctionnelle

$$J = \iint_D \pi(x, y, u_x, u_y)\, dx\, dy, \tag{2.9}$$

où π possède un certain degré de régularité comme dans le problème précédent.

On cherche une solution $u(x, y)$ continue ainsi que ses dérivées partielles d'ordre un et deux, qui a une valeur donnée sur la frontière et rend extrémale la fonctionnelle J. On procède comme plus haut et on aboutit à l'équation d'Euler de la forme

$$\pi_u - \frac{\partial}{\partial x}\pi_{u_x} - \frac{\partial}{\partial y}\pi_{u_y} = 0. \tag{2.10}$$

Les résultats se généralisent évidemment au cas de n variables.

Ainsi, on examine un même problème de la physique mathématique soit avec l'optique des équations différentielles (équations d'Euler), soit en termes de fonctionnelles variationnelles minimisées par une fonction. Dans ce cas la fonction extrémale est solution de l'équation d'Euler.

Soit par exemple le problème

$$\begin{aligned} -\frac{d^2u}{dx^2} + ku &= f, \\ u(0) = u(1) &= 0, \end{aligned} \tag{2.11}$$

où k et f sont suffisamment régulières (k est non négative).

Considérons la fonctionnelle variationnelle

$$J = \int_0^1 \left[\left(\frac{du}{dx}\right)^2 + ku^2 - 2fu\right] dx. \tag{2.12}$$

On cherche l'équation d'Euler correspondante. On pose

$$\pi = \left(\frac{du}{dx}\right)^2 + ku^2 - 2fu.$$

La relation

$$\pi_u - \frac{d}{dx}\pi_{u'} = 0 \tag{2.13}$$

entraîne

$$-\frac{d^2u}{dx^2}+ku=f, \tag{2.14}$$

à laquelle on adjoint la condition

$$u\,(0)=u\,(1)=0. \tag{2.15}$$

Cela signifie que si chaque élément du domaine de définition de J remplit cette condition, l'extrémum de la fonctionnelle (2.12) qu'il réalise, fournit la solution du problème (2.11). L'équivalence des deux positions du problème se démontre de façon rigoureuse par des raisonnements supplémentaires.

Introduisons la notation

$$Au\equiv-\frac{d^2u}{dx^2}+ku$$

et supposons que

$$u\,(0)=u\,(1)=0$$

et que $u\,(x)$ remplit certaines conditions de régularité. L'espace de ces fonctions constitue le domaine $\Phi\,(A)$ de l'opérateur A. Le problème (2.11) s'écrit formellement

$$Au=f,\quad u\in\Phi\,(A).$$

Considérons la fonctionnelle

$$J\,(u)=(Au,\,u)-2\,(f,\,u)=$$

$$=\int_0^1\left(-\frac{d^2u}{dx^2}+ku-2f\right)u\,dx=\int_0^1\left[\left(\frac{du}{dx}\right)^2+ku^2-2fu\right]dx$$

pour $u\in\Phi\,(A)$.

Soit

$$\pi=\left(\frac{du}{dx}\right)^2+ku^2-2fu.$$

On vérifie sans peine que l'équation d'Euler est donnée par le problème

$$Au=f,\quad u\in\Phi\,(A), \tag{2.16}$$

associé au problème variationnel

$$J\,(u)=\min_{v\in\Phi(A)}J\,(v). \tag{2.17}$$

Cherchons les conditions d'équivalence de (2.16) et (2.17). On considère l'équation plus générale

$$Au=f,$$

où A est un opérateur linéaire positif auto-adjoint de domaine $\Phi(A)$, ensemble partout dense dans l'espace de Hilbert H muni du produit scalaire $(,)$, à valeurs dans H, $u \in \Phi(A)$, f étant un élément de H. On a l'affirmation suivante (voir Mikhlin [I]) : si la solution du problème (2.16) existe, elle réalise le minimum de la fonctionnelle

$$J(u) = (Au, u) - 2(u, f).$$

On suppose pour le démontrer que u_0 est solution de (2.1), i.e.

$$Au_0 = f.$$

Soit η un élément non nul arbitraire de $\Phi(A)$ et α un nombre réel quelconque. On définit v_α par l'égalité

$$v_\alpha = u_0 + \alpha\eta,$$

auquel cas

$$J(v_\alpha) = (A(u_0 + \alpha\eta), u_0 + \alpha\eta) - 2(u_0 + \alpha\eta, f).$$

A étant auto-adjoint, on a

$$J(v_\alpha) = J(u_0) + 2\alpha(Au_0 - f, \eta) + \alpha^2(A\eta, \eta).$$

D'où

$$J(v_\alpha) = J(u_0) + \alpha^2(A\eta, \eta).$$

On en tire

$$J(v_\alpha) > J(u_0) \text{ pour tout } \alpha \neq 0 \tag{2.18}$$

vu la positivité de A. Cela signifie que le minimum de la fonctionnelle $J(v)$ est atteint en $v_\alpha = u_0$.

La réciproque est également vraie, à savoir l'élément u de l'espace hilbertien H qui rend minimale la fonctionnelle J et appartient à $\Phi(A)$, est solution de l'équation opératorielle $Au = f$. En effet, soit $u_0 \in \Phi(A)$ minimisant $J(u)$ et η un élément quelconque du même domaine. Il est connu que $w = \alpha u + \beta v$ (α et β étant des constantes) se trouve pour tout u, $v \in \Phi(A)$ dans $\Phi(A)$. Aussi $v_\alpha = u_0 + \alpha\eta \in \Phi(A)$. Sous l'hypothèse de u_0 réalisant le minimum de J, on a

$$J(u_0 + \alpha\eta) \geqslant J(u_0). \tag{2.19}$$

Admettons que α est réel. Avec A supposé symétrique, la relation (2.19) conduit à

$$2\alpha(Au_0 - f, \eta) + \alpha^2(A\eta, \eta) \geqslant 0,$$

ce qui ne peut avoir lieu que si

$$(Au_0 - f, \eta) = 0. \tag{2.20}$$

Ainsi, $Au_0 = f$ est orthogonal à tous les éléments de $\Phi(A)$, si bien que

$$Au_0 - f = 0.$$

Plus bas, nous parlerons en détail de la position variationnelle de problèmes de la physique mathématique et décrirons des méthodes principales de résolution. La plupart des résultats sont empruntés à Mikhlin [I] où l'on trouve de même la démonstration des théorèmes énoncés.

Nous allons raisonner sur l'équation différentielle elliptique

$$Lu \equiv -\sum_{i,\,j=1}^{2} \frac{\partial}{\partial x_i} A_{ij}(x) \frac{\partial u}{\partial x_j} + \sum_{i=1}^{2} B_i(x) \frac{\partial u}{\partial x_i} + q(x)\, u = f, \qquad x = (x_1, x_2) \in D, \tag{2.21}$$

donnée dans un domaine borné D avec la condition aux limites de la forme

$$u = 0, \qquad x \in \partial D \tag{2.22}$$

(problème de Dirichlet).

On suppose que l'opérateur

$$L_0 = -\sum_{i,\,j=1}^{2} \frac{\partial}{\partial x_i} A_{ij}(x) \frac{\partial}{\partial x_j} \tag{2.23}$$

agit dans l'espace de Hilbert $F = L_2(D)$, a pour domaine l'ensemble $\Phi(L_0)$ de fonctions $u \in L_2(\Omega)$ telles qu'on ait (2.22) et $Lu \in L_2(\Omega)$, est auto-adjoint au sens de Lagrange et n'est pas dégénéré, i.e. tout vecteur non nul $\xi = (\xi_1, \xi_2)$ satisfait à l'inégalité

$$\inf_{x \in D} \sum_{i,\,j=1}^{2} A_{ij}(x)\, \xi_i \xi_j \geqslant \mu_0 \sum_{i=1}^{2} \xi_i^2 \tag{2.24}$$

avec une constante positive μ_0. On admet de plus que la fonction $q(x)$ est non négative dans D et que le problème (2.21), (2.22) possède une solution, d'où des questions : quel est le degré de régularité des données ? est-ce qu'il s'agit d'une solution classique ou d'une solution généralisée ? et ainsi de suite.

Nous supposerons, *primo*, que la solution est classique et appartient à l'espace de Sobolev $\overset{\circ}{W}{}_2^1$ formé de fonctions $\in L_2(D)$ qui présentent dans D des dérivées généralisées de carré sommable et s'annulent sur la frontière ∂D. On met sur $\overset{\circ}{W}{}_2^1$ la norme (voir n° 1.1)

$$\|u\|_{\overset{\circ}{W}{}_2^1} = \left\{ \int_D u^2\, dD + \int_D \left[\left(\frac{\partial u}{\partial x_1} \right)^2 + \left(\frac{\partial u}{\partial x_2} \right)^2 \right] dD \right\}^{1/2}. \tag{2.25}$$

On admet que les propriétés des données du problème (2.21), (2.22) (par exemple la régularité des coefficients et de la frontière) garantissent l'appartenance de la solution à l'espace donné. Nous suppose-

rons, *secundo*, si besoin est, que les solutions ou les données remplissent des exigences supplémentaires nécessaires de régularité. Avec ces hypothèses, on se débrouille mieux dans les principes de base de la discrétisation d'équations aux dérivées partielles, qui constitue le but principal de ce chapitre.

2.2.1. Méthode de Ritz

Admettons que B_i, $i = 1, 2$, sont identiquement nulles dans D. Avec cette hypothèse supplémentaire, on prouve facilement l'identité intégrale

$$(L\varphi, \psi) \equiv \int_D \psi L\varphi \, dD = \int_D \left\{ \sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial \varphi}{\partial x_i} \frac{\partial \psi}{\partial x_j} + q(x) \varphi\psi \right\} dD \quad (2.26)$$

pour toutes les fonctions φ et ψ du domaine de L.

Associons au problème initial (2.21), (2.22) le problème variationnel de trouver $u \in \Phi(L)$, $\Phi(L)$ étant le domaine de L, qui réalise le minimum de la fonctionnelle quadratique

$$J(u) = \int_D uLu \, dD - 2 \int_D fu \, dD.$$

Selon (2.26), le domaine de définition $\Phi(J)$ de la fonctionnelle J contient le domaine de L, i.e. $\Phi(L) \subseteq \Phi(J)$. Munissons $\Phi(J)$ du nouveau produit scalaire défini par

$$(\varphi, \psi)_L = (L\varphi, \psi) = \int_D \left\{ \sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial \varphi}{\partial x_i} \frac{\partial \psi}{\partial x_j} + q(x) \varphi\psi \right\} dD$$

et de la norme associée

$$\| \varphi \|_L = (\varphi, \varphi)_L^{1/2}.$$

En complétant $\Phi(J)$ pour la norme introduite, on obtient l'espace normé complet F_L, espace énergétique engendré par L dont chaque élément appartient à F.

Formulons le problème variationnel proposé en termes de F_L. On dit que si $u \in \Phi(L)$ est la solution du problème (2.21), (2.22), elle minimise J dans F_L. Inversement, si

$$J(u) = \min_{v \in F_L} J(v) \quad (2.27)$$

et $u \in \Phi(L)$, alors u est la solution de (2.21), (2.22).

Comme il y a unicité, on considère naturellement, au lieu de (2.21), (2.22), le problème variationnel (2.27) et les méthodes de résolution correspondantes.

Parmi celles-ci, la méthode de Ritz est la plus connue. Appliquons-la à l'équation opératorielle

$$Lu = f \tag{2.28}$$

dans l'espace hilbertien F à produit scalaire (u, v), l'opérateur L étant auto-adjoint et défini positif dans F. Nous savons déjà que la recherche de la solution de (2.28) équivaut au problème de trouver un élément de F qui réalise le minimum de la fonctionnelle

$$J(u) = (Lu, u) - 2(u, f). \tag{2.29}$$

Introduisons une suite des sous-espaces $F_h \subseteq F_L$ de dimension n définis par une suite infinie de paramètres $h_1, h_2, \ldots, h_k \xrightarrow[k\to\infty]{} 0$. Nous dirons que la suite $\{F_h\}$ est complète dans F_L s'il existe pour tous les $u \in F_L$ et $\varepsilon > 0$ un $\hat{h} = \hat{h}(u, \varepsilon) > 0$ tel que

$$\inf_{w \in F_h} \|u - w\| < \varepsilon \tag{2.30}$$

quels que soient $h < \hat{h}$. Autrement dit, la propriété de complétude de $\{F_h\}$ signifie que tout élément $u \in F$ est approché aussi exactement qu'on le veut par les éléments des F_h.

La méthode de Ritz consiste donc à chercher un élément $u^h \in F_h$ qui minimise $J(u)$ dans F_h. On affirme qu'avec L défini positif dans F_h, si la suite $\{F_h\}$ est complète dans F_L, alors la suite d'approximations $\{u^h\}$ de Ritz converge dans F_L vers la solution u de l'équation (2.28).

Quand la base de F_h est supposée connue et formée de fonctions $\{\varphi_i^h\}_{i=1}^{N_h}$, le problème de trouver $u^h \in F_h$ équivaut à la recherche des coefficients $\{\alpha_i\}_{i=1}^{N_h}$ du développement

$$u^h = \sum_{i=1}^{N_h} \alpha_i \varphi_i^h \tag{2.31}$$

en minimisation de J. On substitue comme d'habitude (2.31) dans la fonctionnelle J et on égale à 0 les dérivées $\frac{\partial J(u^h)}{\partial \alpha_i}$, $i = 1, \ldots, N_h$. Le système d'équations algébriques linéaires obtenu s'écrit

$$A\alpha = g, \tag{2.32}$$

avec α et g des vecteurs de dimension N_h,

$$g_i = (f, \varphi_i^h), \tag{2.33}$$

et $A = (a_{ij})$ est la matrice de Gram du produit scalaire du système de vecteurs $\{\varphi_i\}$ avec l'opérateur L, i.e.

$$a_{ij} = (L\varphi_i^h, \varphi_j^h), \qquad 1 \leqslant i, j \leqslant N_h. \tag{2.34}$$

Comme $a_{ij} = a_{ji}$, la matrice A est symétrique. L'inégalité

$$(A\xi, \xi) = \left(L\left(\sum_{i=1}^{N_h} \xi_i \varphi_i^h\right), \sum_{i=1}^{N_h} \xi_i \varphi_i^h\right) > 0 \tag{2.35}$$

avec $\xi \neq 0$ entraîne sa définie positivité.

Si la solution du problème variationnel est limite d'une suite d'éléments qui ne vérifient pas tous les conditions aux limites du problème différentiel, ces conditions sont dites naturelles (par exemple, les conditions de Neumann sont naturelles). Si les conditions aux limites sont remplies par chaque terme de la suite, elles s'appellent parfois conditions essentielles (la condition de Dirichlet est du nombre). Lions [I] et Babuška [V] ont proposé une méthode de pénalisation qui permet de ramener de façon approchée un problème avec conditions aux limites essentielles à un problème variationnel avec conditions naturelles. S'agissant du problème de Dirichlet, on pose par exemple le problème de minimisation de la fonctionnelle

$$J_\varepsilon^f(u_\varepsilon) = \int_D (u_\varepsilon L u_\varepsilon)\, dD - 2\int_D f u_\varepsilon\, dD + \frac{1}{\varepsilon}\int_{\partial D} u_\varepsilon^2\, ds,$$

avec ε un paramètre suffisamment petit.

Quand le problème de Dirichlet discrétisé se rapporte à l'équation de Poisson (voir Babuška [V]), on choisit $\varepsilon = h^\sigma$, $\sigma > 0$, auquel cas $f \in W_2^k$ s'évalue par $\| u - u_\varepsilon \|_{W_2^1} \leqslant ch^\mu \| f \|_{W_2^k}$, et

$$\mu = \min\left(\frac{\sigma}{2},\ k - 1 - \frac{\sigma}{2}\right).$$

Cette formule donne pour k donné un σ tel que l'ordre de convergence μ soit maximal.

2.2.2. Méthode de Galerkin

L'inconvénient majeur de la technique de Ritz est qu'elle joue uniquement pour les équations à opérateurs définis positifs autoadjoints. Ce défaut n'est pas propre à cet autre procédé variationnel qu'est la méthode de Galerkin (ou de Boubnov-Galerkin). Décrivons-la pour l'équation opératorielle

$$Lu = f$$

dans l'espace hilbertien F ($\Phi(L)$ est dense dans F), où $L = L_0 + K$, L_0 étant un opérateur symétrique défini positif et $\Phi(L_0) \subseteq \Phi(K)$.

On introduit une fois de plus une suite des sous-espaces $F_h \subset F_{L_0}$, $h = h_1, h_2, \ldots$, de dimension finie ayant les bases $\{\varphi_i^h\}_{i=1}^{N_h}$. L'approximation est cherchée sous la forme

$$u^h = \sum_{i=1}^{N_h} \alpha_i \varphi_i^h, \tag{2.36}$$

avec α_i choisis de façon que le résidu $Lu^h - f$ soit orthogonal à tous les éléments de F_h. On aboutit au système

$$A\alpha = g, \tag{2.37}$$

où

$$\alpha_{ij} = (L\varphi_j^h, \varphi_i^h), \quad g_i = (f, \varphi_i^h), \qquad i, j = 1, \ldots, N_h. \tag{2.38}$$

On calcule les coefficients $\{\alpha_i\}$ et on obtient sans peine la solution approchée (2.36). On vérifie facilement que le système (2.37) coïncide pour L auto-adjoint positif avec le système (2.32) donné par le procédé de Ritz.

Voyons la convergence des solutions approchées (2.36). On a le résultat suivant.

Si la solution u du problème $Lu = f$ existe et est unique dans F, la suite $\{F_h\}$ est complète dans F (au sens de la définition du point 2.2.1) et l'opérateur $L_0^{-1}K$ compact dans F, alors les approximations successives u^h de la méthode de Galerkin convergent dans F vers la solution exacte u.

On montre que ces conditions sont satisfaites pour le problème modèle (2.21), (2.22) si l'on prend pour F l'espace $\mathring{W}_2^1$. Dans la méthode étudiée on ne suppose plus les fonctions $\{B_i(x)\}_{i=1}^2$ identiquement nulles.

Voici une modification de la technique de Galerkin avec L_0 un opérateur qui n'est en général pas symétrique et défini positif. Soit L_0^{-1} borné défini sur F tout entier. L'équation considérée équivaut à

$$u + L_0^{-1}Ku = f', \quad f' = L_0^{-1}f. \tag{2.39}$$

Désignons par F_1 l'espace de Hilbert muni du produit scalaire $(u, v)_1 = (L_0u, L_0v)$ et de la norme $\| u \|_1 = \| L_0u \|$. La méthode de Galerkin s'énonce pour (2.39) de façon suivante. Soit F_h des sous-espaces de F_1 de dimension finie de bases $\{\varphi_i^h\}_{i=1}^{N_h}$. On cherche la solution approchée sous la forme $u^h = \sum_{i=1}^{N_h} \alpha_i \varphi_i^h$, les inconnues $\{\alpha_i\}_{i=1}^{N_h}$ étant définies par le système linéaire

$$(u^h, \varphi_i^h)_1 + (L_0^{-1}Ku^h, \varphi_i^h)_1 = (f', \varphi_i^h)_1, \qquad i = 1, \ldots, N_h. \tag{2.40}$$

Il est déjà apparu que s'il y a unicité pour $Lu = f$, la suite $\{F_h\}$ est complète dans F_1 et l'opérateur $L_0^{-1}K$ compact dans F_1, alors la suite des u^h converge vers la solution exacte dans F aussi bien que dans F_1.

Remarque. Le système (2.40) équivaut à

$$(L_0u^h, L_0\varphi_i^h) + (Ku^h, L_0\varphi_i^h) = (f, L_0\varphi_i^h), \qquad i = 1, \ldots, N_h \tag{2.41}$$

qui définit la méthode des moments, un autre procédé variationnel de grand emploi.

Dans le procédé (2.40) (ou (2.41)), on se heurte entre autres au problème de choisir les fonctions de base. Si $\{\varphi_i^h\}_{i=1}^{N_h}$ possédant certaines propriétés d'approximation sont données à priori, il arrive que le système $\{\psi_i^h\}_{i=1}^{N_h}$, $\psi_i^h = L_0\varphi_i^h$, se prête mal à l'étude, ce qui se répercute de façon fâcheuse sur l'évaluation de la vitesse de convergence, l'analyse des singularités de la solution et des caractères spéciaux au problème.

Considérons un algorithme de génération des fonctions de base qui a été utilisé dans plusieurs problèmes aux limites par Marchouk et Agochkov [V ; 5].

L'image de l'opérateur K et la fonction f sont supposées dans un sous-espace $F(K, f) \subset F$. On se donne dans $F(K, f)$ un système de fonctions de base $\{\psi_i^h\}_{i=1}^{N_h}$ à support borné d'ordre h telles que la suite $\{\psi_i^h\}$, $h = h_1, h_2, \ldots$, soit complète dans $F(K, f)$. Formons les fonctions $\{\varphi_i^h\}_{i=1}^{N_h}$, avec $\varphi_i^h = L_0^{-1}\psi_i^h$. Elles sont linéairement indépendantes pour tout h et jouent le rôle de fonctions de base dans la résolution de (2.39) par la méthode de Galerkin.

Précisons certaines propriétés de l'algorithme décrit.

Les fonctions $\{\varphi_i^h\}_{i=1}^{N_h}$ possèdent par construction les singularités de u propres à L_0 et, avec $\{\psi_i^h\}_{i=1}^{N_h}$ choisi de façon spéciale, on tient compte des singularités connues à priori de $w = f - Ku$.

Il arrive que w présente des propriétés de dérivabilité meilleures que la solution même. Dans ce cas, on essaie de bien approcher w par des fonctions de base initiales en nombre insignifiant et on escompte la convergence rapide $u^h \to u$.

Si la solution de (2.39) dépend des variables x_i, $i = 1, \ldots, n$ et tous les éléments de $F(K, f)$ sont fonction de x_i, $i = 1, \ldots, m < n$, seule, il suffit évidemment que $\{\psi_i^h\}_{i=1}^{N_h}$ utilisées dépendent seulement de x_i, $i = 1, \ldots, m < n$, et on approche avec $w \in F(K, f)$. Quant à la solution u, elle est approchée par u^h par rapport à toutes les variables. Cela diminue considérablement dans la pratique le nombre de fonctions de base nécessaires et, partant, l'ordre du système (2.40), chose importante lorsqu'on a affaire aux grands problèmes de la physique mathématique.

Récrivons (2.40) de façon équivalente :

$$\sum_{i=1}^{N_h} \alpha_i(\psi_i^h, \psi_j^h) = -\sum_{i=1}^{N_h} \alpha_i(K\varphi_i^h, \psi_j^h) + (f, \psi_j^h), \qquad (2.42)$$

$$j = 1, \ldots, N_h.$$

On constate sans peine que ψ_j^h, $j = 1, \ldots, N_h$, étant à support borné, il apparaît dans le premier membre de (2.42) une matrice bande (ou creuse), ce qui facilite des fois la résolution par des méthodes itératives. La propriété indiquée de ψ_j^h, $j = 1, \ldots, N_h$, a également

pour effet de simplifier le calcul des $\{\varphi_i^h\}_{i=1}^{N_h}$, des éléments des matrices et des valeurs de (f, ψ_j^h).

Il y a lieu de supposer que l'utilisation des bases spéciales $\{\varphi_i^h\}$ assure une efficacité suffisante de la méthode de Galerkin pour des problèmes aux limites où L_0^{-1} est construit assez vite (par exemple pour L_0 inversible sous forme explicite: opérateur différentiel dans l'équation de transport, laplacien dans un carré, un disque, etc.).

2.2.3. Méthode des moindres carrés

La méthode des moindres carrés est très employée dans les problèmes aux limites de la physique mathématique. Décrivons-la pour l'équation opératorielle

$$Lu = f \tag{2.43}$$

dans l'espace hilbertien F.

Soit F_h des sous-espaces n-dimensionnels de F de bases $\varphi_1^h, \ldots, \varphi_{N_h}^h$, $F_h \subseteq \Phi(L)$, domaine de L (on peut, au lieu de L, considérer son prolongement $\hat{L}$ tel que $\Phi(\hat{L})$ soit complet).

On forme les solutions approchées (2.36) au sens des moindres carrés par les égalités

$$\frac{\partial}{\partial \alpha_i} \| Lu - f \| = 0, \qquad i = 1, \ldots, N_h, \tag{2.44}$$

et il vient le système linéaire (2.37) de matrice $A = (a_{ij})$ et de vecteur $g = (g_i)$, où

$$a_{ij} = (L\varphi_i^h, L\varphi_j^h), \quad g_i = (f, L\varphi_i^h), \qquad 1 \leqslant i, j \leqslant N_h. \tag{2.45}$$

Si L est un opérateur non singulier dans $\Phi(L)$, il y a évidemment symétrie et définie positivité de la matrice A.

On formule des conditions suffisantes de convergence de la méthode des moindres carrés.

Les approximations successives u^h convergent dans F vers la solution exacte u de (2.43) si celle-ci admet une solution unique, la suite LF_h est complète dans $\Phi(L)$ et il y a existence et borne de l'opérateur L^{-1}.

La deuxième condition s'interprète comme suit. Comme au numéro 2.2.1, la complétude de LF_h signifie *) qu'on indique pour tout $u \in \Phi(L)$ et tout $\varepsilon > 0$ un $\hat{h} = \hat{h}(u, \varepsilon) > 0$ tel que

$$\inf_{\omega \in F_h} \| Lu - L\omega \| < \varepsilon \tag{2.46}$$

*) L'écriture LF_h a un sens car $F_h \subset \Phi(L)$ par hypothèse.

quels que soient F_h, $h < \hat{h}$. Il est évident que sous l'hypothèse de u (solution de (2.43)) unique, les conditions du théorème garantissent aussi bien $Lu^h \underset{h\to 0}{\longrightarrow} Lu$ que $u^h \underset{h\to 0}{\longrightarrow} u$.

S'agissant des problèmes aux limites, la méthode des moindres carrés le cède aux deux techniques précédentes pour ce qui est de vérifier si la solution limite remplit les conditions aux limites. Voici deux voies possibles à suivre.

1) On impose aux fonctions des espaces F_h de vérifier exactement les conditions aux limites. Cette tactique semble tout indiquée, mais elle s'avère difficile à réaliser numériquement même pour le problème mixte relatif à l'équation elliptique (2.21).

2) On peut utiliser la « méthode des poids » et poser un problème variationnel auxiliaire (voir par exemple Bramble et Schatz [V]). On procède comme suit. On associe à l'équation aux dérivées partielles d'ordre $2m$

$$Lu = f \quad \text{dans} \quad D \tag{2.47}$$

avec les conditions aux limites

$$L_i u = f_i \quad \text{sur } \partial D, \qquad i = 1, \ldots, m, \tag{2.48}$$

la fonctionnelle

$$J_h(u) = \| Lu - f \|^2 + \sum_{i=1}^{m} c_i(h) \| L_i u - f_i \|^2, \tag{2.49}$$

avec $\{c_i(h)\}_{i=1}^m$ des fonctions positives du paramètre h caractéristique de la suite des F_h. On a pour l'analogue aux différences du problème auto-adjoint (2.47), (2.48) par rapport aux solutions régulières: $c_i(h) = h^{-2(2m-m_i-1/2)}$, m_i étant l'ordre de dérivation le plus élevé dans L_i. On cherche les approximations u^h par les moindres carrés comme solutions des problèmes variationnels

$$\inf_{u \in F_h} J_h(u) = J(u^h).$$

Les fonctions u^h convergent vers u quand $h \to 0$, et l'équation (2.47) est asymptotiquement vérifiée, ainsi que les conditions (2.48). Ceci étant, les éléments des sous-espaces F_h ne vérifient pas nécessairement les conditions aux limites.

2.3. Schémas variationnels aux différences pour les équations à coefficients discontinus

Les travaux de Lions [I] [II] [V], Oganessian, Rivkind, Roukhovetz [V ; 8], Aubin [V], Birkhoff, Schultz, Varga [V], Babuška [V] ont ouvert une ère de la construction des équations aux différences à la lumière des principes du Calcul des Variations. Plusieurs idées intéressantes sont nées dernièrement dont la plus prometteuse est

d'utiliser les fonctions à support borné en qualité de fonctions d'essai (ces fonctions sont différentes de 0 dans une région relativement restreinte (de l'ordre du pas du maillage) et identiquement nulles partout ailleurs). On a vu l'avantage qu'on a parfois de chercher la solution sous forme de combinaison linéaire de fonctions à support borné, les coefficients inconnus de la combinaison étant choisis à partir de la condition de minimum d'une fonctionnelle variationnelle. On a procédé de la sorte pour différentes classes de problèmes et abouti à un algorithme de discrétisation très efficace. Nous allons illustrer cet algorithme sur un problème relatif à la diffusion de substance en unidimensionnel.

2.3.1. Construction d'équations de diffusion discrétisées simples par la méthode de Ritz

On sait que la méthode de Ritz joue seulement pour les problèmes à opérateurs auto-adjoints. On cherche la solution approchée du problème aux limites auto-adjoint (1.1), (1.2) et on introduit la fonctionnelle variationnelle

$$J = \int_0^1 \left[p \left(\frac{d\varphi}{dx} \right)^2 + q\varphi^2 - 2f\varphi \right] dx. \tag{3.1}$$

Comme cette fonctionnelle atteint son minimum sur la solution du problème (1.1), (1.2) du n° 2.1 (voir n° 2.2), on construit la solution approchée sur le réseau D_h de façon que ses paramètres libres soient choisis en minimisation de (3.1). On cherche donc la solution approchée sous forme de fonction linéaire par morceaux continue de $\mathring{W}_2^1$

$$\varphi^h(x) = \frac{x_{k+1} - x}{\Delta x_{k+1/2}} \varphi_k + \frac{x - x_k}{\Delta x_{k+1/2}} \varphi_{k+1}, \quad x_k \leqslant x \leqslant x_{k+1}, \tag{3.2}$$

où $\Delta x_{k+1/2} = x_{k+1} - x_k$, φ_k sont les valeurs approchées de la solution $\varphi(x)$ aux nœuds. Ces valeurs sont certes inconnues, si bien que le problème consiste à leur recherche par minimisation de la fonctionnelle $J(\varphi)$.

Introduisons sur l'intervalle $x_k \leqslant x \leqslant x_{k+1}$ deux fonctions linéaires $\omega_1(x)$ et $\omega_2(x)$ définies par

$$\omega_1(x) = \frac{x - x_k}{\Delta x_{k+1/2}}, \qquad \omega_2(x) = \frac{x_{k+1} - x}{\Delta x_{k+1/2}}. \tag{3.3}$$

La formule d'interpolation (3.2) se récrit

$$\varphi^h(x) = \omega_1(x)\varphi_{k+1} + \omega_2(x)\varphi_k. \tag{3.4}$$

Les fonctions $\omega_1(x)$, $\omega_2(x)$ dépendent en général de leur intervalle de définition et on ferait mieux d'utiliser les notations

$\omega_{1,\,k+1/2}(x)$, $\omega_{2,\,k+1/2}(x)$. Dans la suite nous omettrons partout les indices $k+1/2$ parce que l'appartenance en question ne suscitera jamais de doute. Les renseignements à priori sur la solution de (1.1), (1.2) nous font construire la solution approchée $\varphi^h(x)$ de façon qu'elle soit continue sur l'intervalle $0 \leqslant x \leqslant 1$. Les fonctions $p(x) \geqslant p_0 > 0$, $q(x) \geqslant 0$ et $f(x)$ sont supposées continues par morceaux avec des discontinuités de première espèce. On admet x_k pour points de discontinuité. L'écriture commode de la fonctionnelle minimisante $J(\varphi^h)$ est

$$J(\varphi^h) = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \left[p \left(\frac{d\varphi}{dx} \right)^2 + q\varphi^2 - 2f\varphi \right] dx. \tag{3.5}$$

Portons (3.4) dans (3.5) et prenons en considération les relations

$$\frac{d\omega_1}{dx} = \frac{1}{\Delta x_{k+1/2}}, \quad \frac{d\omega_2}{dx} = -\frac{1}{\Delta x_{k+1/2}}, \quad x_k < x < x_{k+1},$$

il vient

$$J(\varphi^h) = \sum_{k=0}^{n-1} \int_{x_k}^{x_k+1} \left[p \frac{\varphi_{k+1}^2 - 2\varphi_{k+1}\varphi_k + \varphi_k^2}{\Delta x_{k+1/2}^2} + \right.$$

$$\left. + q(\omega_1^2\varphi_{k+1}^2 + 2\omega_1\omega_2\varphi_{k+1}\varphi_k + \omega_2^2\varphi_k^2) - 2f(\omega_1\varphi_{k+1} + \omega_2\varphi_k) \right] dx. \tag{3.6}$$

On choisit φ_k, $k = 1, 2, \ldots, n-1$, qui minimisent la fonctionnelle $J(\varphi^h)$. Le minimum a lieu sous la condition

$$\frac{\partial J}{\partial \varphi_k} = 0; \qquad k = 1, \ldots, n-1. \tag{3.7}$$

On dérive (3.6) par rapport à φ_k et on égale le résultat à zéro :

$$\frac{p_{k-1/2}}{\Delta x_{k-1/2}}(\varphi_k - \varphi_{k-1}) - \frac{p_{k+1/2}}{\Delta x_{k+1/2}}(\varphi_{k+1} - \varphi_k) + q_{k+1/2}^{1,2}\varphi_{k+1} +$$

$$+ (q_{k+1/2}^{2,2} + q_{k-1/2}^{1,1})\varphi_k + q_{k-1/2}^{1,2}\varphi_{k-1} = F_k, \tag{3.8}$$

où

$$p_{k+1/2} = \frac{1}{\Delta x_{k+1/2}} \int_{x_k}^{x_{k+1}} p\,dx; \qquad q_{k+1/2}^{i,j} = \int_{x_k}^{x_{k+1}} \omega_i\omega_j q\,dx;$$

$$F_k = \int_{x_{k-1}}^{x_k} f\omega_1\,dx + \int_{x_k}^{x_{k+1}} f\omega_2\,dx.$$

L'équation aux différences (3.8) plus les conditions aux limites

$$\varphi_0 = 0, \quad \varphi_n = 0$$

définissent complètement le problème aux limites.

L'équation (3.8) et les conditions ci-dessus se mettent sous forme de problème aux différences à trois points

$$\begin{gathered} a_k\varphi_{k-1} - b_k\varphi_k + c_k\varphi_{k+1} = -F_k, \\ \varphi_0 = 0, \quad \varphi_n = 0, \end{gathered} \tag{3.9}$$

où

$$\begin{gathered} a_k = \frac{p_{k-1/2}}{\Delta x_{k-1/2}} - q^{1,2}_{k-1/2}; \quad c_k = \frac{p_{k+1/2}}{\Delta x_{k+1/2}} - q^{1,2}_{k+1/2}; \\ b_k = \frac{p_{k+1/2}}{\Delta x_{k+1/2}} + \frac{p_{k-1/2}}{\Delta x_{k-1/2}} + (q^{1,1}_{k+1/2} + q^{2,2}_{k-1/2}). \end{gathered} \tag{3.10}$$

Plaçons-nous dans le cas particulier où p et q sont constantes sur chaque intervalle $x_k < x < x_{k+1}$. On vérifie de suite que

$$a_k = \frac{p_{k-1/2}}{\Delta x_{k-1/2}} - \frac{\Delta x_{k-1/2}}{6} q_{k-1/2}; \quad c_k = \frac{p_{k+1/2}}{\Delta x_{k+1/2}} - \frac{\Delta x_{k+1/2}}{6} q_{k+1/2}; \tag{3.11}$$

$$b_k = \frac{p_{k+1/2}}{\Delta x_{k+1/2}} + \frac{p_{k-1/2}}{\Delta x_{k-1/2}} + \frac{1}{3}(\Delta x_{k+1/2} q_{k+1/2} + \Delta x_{k-1/2}\, q_{k-1/2}).$$

Ici

$$p_{k+1/2} = p(x_{k+1/2}); \qquad q_{k+1/2} = q(x_{k+1/2}).$$

Supposons connue la solution de (3.9) et reconstituons la solution continue $\varphi^h(x)$ à l'aide des formules (3.4). Le grand avantage de l'approche variationnelle est de construire à priori la solution approchée par un polynôme d'interpolation et de chercher la solution stricte sous forme de fonction linéaire par morceaux continue qui garantit par construction la meilleure approximation dans la classe de fonctions linéaires par morceaux continues.

Avec la méthode variationnelle, on approche la solution aussi exactement qu'on le veut. A cet effet, on remplace les formules d'interpolation simples (3.4) par d'autres plus précises. On utilise de même de façon formelle les polynômes d'interpolation de Lagrange.

Procédons à l'étude de la propriété d'approximation des équations aux différences obtenues. L'équation (3.8) se récrit

$$(A^h\varphi)_k = -\frac{1}{\Delta x_k}\left\{p_{k+1/2}\frac{\varphi_{k+1}-\varphi_k}{\Delta x_{k+1/2}} - p_{k-1/2}\frac{\varphi_k-\varphi_{k-1}}{\Delta x_{k-1/2}} - \right.$$

$$\left. - [q^{1,2}_{k+1/2}\varphi_{k+1} + (q^{2,2}_{k+1/2} + q^{1,1}_{k-1/2})\varphi_k + q^{1,2}_{k-1/2}\varphi_{k-1}]\right\} = f^h_k, \tag{3.12}$$

où

$$f^h_k = \frac{1}{\Delta x_k} F_k, \tag{3.13}$$

avec $\Delta x_k = \frac{1}{2}(\Delta x_{k+1/2} + \Delta x_{k-1})$. Etant données la continuité et la grande régularité de la solution $\varphi(x)$ et de tous les p, q, f sur les intervalles $x_k < x < x_{k+1}$, on montre grâce à la formule de Taylor que

$$\|A^h(\varphi)_h - f^h\| \leqslant Mh^\alpha,$$

avec α défini par les propriétés du réseau (voir n° 2.1). Les développements suivants coïncident avec ceux du n° 2.1.

2.3.2. Construction de schémas aux différences simples par la méthode de Galerkin (des éléments finis)

Soit le problème (1.1), (1.2) défini dans $0 \leqslant x \leqslant 1$. Recouvrons ce domaine par un système d'intervalles $x_{k-1} \leqslant x \leqslant x_k$ et associons à chaque $k \geqslant 1$ une fonction de $\mathring{W}_2^1$

$$\omega_k(x) = \begin{cases} 0, & 0 \leqslant x \leqslant x_{k-1}, \\ \omega_1(x), & x_{k-1} \leqslant x \leqslant x_k, \\ \omega_2(x), & x_k \leqslant x \leqslant x_{k+1}, \\ 0, & x_{k+1} \leqslant x \leqslant x_N = 1. \end{cases} \tag{3.14}$$

Il ressort de la fig. 1 que $\omega_k(x)$, $k = 1, \ldots, N-1$, ont pour domaine de définition tout le segment $0 \leqslant x \leqslant 1$. Ce sont des fonctions continues qui s'annulent partout en dehors de $x_{k-1} < x < x_{k+1}$ où elles sont formées de deux morceaux linéaires. Leur maximum égal à 1 est réalisé au point $x = x_k$. Ces fonctions possèdent une propriété analogue dans un sens à celle d'être complet: leur système est complet au sens que toute fonction $\varphi(x)$ continue linéaire par parties qui présente peut-être des inflexions aux nœuds $\{x_k\}$, est représentée par une combinaison linéaire de ces fonctions, i.e.

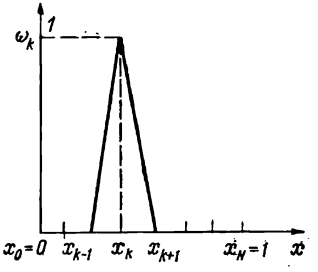


Fig. 1

$$\varphi(x) = \sum_k \varphi_k \omega_k(x), \tag{3.15}$$

les coefficients de Fourier φ_k étant les valeurs de $\varphi(x)$ aux points x_k.

Les fonctions $\omega_k(x)$ jouissent de plus d'une sorte d'orthogonalité. Si l'on définit le produit scalaire par la relation

$$(g, h) = \int_0^1 gh\,dx,$$

alors on a pour $\omega_k(x)$

$$\int_0^1 \omega_k(x)\,\omega_n(x)\,dx = \begin{cases} 0, & n \leqslant k-2, \\ \frac{1}{6}\Delta x_{k-1/2}, & n = k-1, \\ \frac{1}{3}(\Delta x_{k-1/2} + \Delta x_{k+1/2}), & n = k, \\ \frac{1}{6}\Delta x_{k+1/2}, & n = k+1, \\ 0, & n \geqslant k+2, \end{cases} \tag{3.16}$$

d'où leur orthogonalité à toutes les $\omega_n(x)$, sauf ω_{k-1}, ω_k et ω_{k+1}, ce qui constitue le caractère spécifique de la base choisie. On forme avec $\omega_k(x)$ définies les équations aux différences finies. A cet effet, on considère l'équation de diffusion

$$\begin{gathered} -\frac{d}{dx}\,p\,\frac{d\varphi}{dx} + q\varphi = f \text{ dans } D, \\ \varphi(0) = 0, \quad \varphi(1) = 0. \end{gathered} \tag{3.17}$$

Selon le schéma de Galerkin, on multiplie scalairement par ω_n, il vient

$$\int_0^1 \left(-\frac{d}{dx}\,p\,\frac{d\varphi}{dx} + q\varphi - f\right)\omega_n(x)\,dx = 0. \tag{3.18}$$

Transformons la fonctionnelle (3.18):

$$\int_0^1 \left[p\,\frac{d\varphi}{dx}\cdot\frac{d\omega_n}{dx} + (q\varphi - f)\,\omega_n\right]dx = 0. \tag{3.19}$$

L'intégration par parties s'est faite compte tenu de $\omega_n(0) = \omega_n(1) = 0$. On met (3.19) sous forme de somme

$$\sum_k \int_{x_{k-1}}^{x_{k+1}} \left[p\,\frac{d\varphi}{dx}\cdot\frac{d\omega_n}{dx} + (q\varphi - f)\,\omega_n\right]dx = 0. \tag{3.20}$$

Etant donnée la forme de $\omega_n(x)$, on trouve sans peine

$$\begin{aligned} \int_{x_{k-1}}^{x_k} p\,\frac{d\varphi}{dx}\cdot\frac{d\omega_n}{dx}\,dx &= \begin{cases} 0 \text{ si } n \neq k, \\ \frac{p_{k-1/2}}{\Delta x_{k-1/2}}(\varphi_k - \varphi_{k-1}) \text{ si } n = k; \end{cases} \\ \int_{x_k}^{x_{k+1}} p\,\frac{d\varphi}{dx}\cdot\frac{d\omega_n}{dx}\,dx &= \begin{cases} 0 \text{ si } n \neq k, \\ -\frac{p_{k+1/2}}{\Delta x_{k+1/2}}(\varphi_{k+1} - \varphi_k) \text{ si } n = k; \end{cases} \end{aligned} \tag{3.21}$$

и

$$\int_{x_{k-1}}^{x_k} q\varphi\omega_k\,dx = q^{1,2}_{k-1/2}\varphi_{k-1} + q^{1,1}_{k-1/2}\varphi_k,$$
$$\int_{x_k}^{x_{k+1}} q\varphi\omega_k\,dx = q^{2,2}_{k+1/2}\varphi_k + q^{1,2}_{k+1/2}\varphi_{k+1}. \qquad (3.22)$$

(Les notations sont en conformité avec (3.8).)

Portons (3.21), (3.22) dans (3.20), il vient l'équation aux différences

$$\frac{p_{k-1/2}}{\Delta x_{k-1/2}}(\varphi_k - \varphi_{k-1}) - \frac{p_{k+1/2}}{\Delta x_{k+1/2}}(\varphi_{k+1} - \varphi_k) + \\ + q^{1,2}_{k-1/2}\varphi_{k-1} + (q^{1,1}_{k-1/2} + q^{2,2}_{k+1/2})\varphi_k + q^{1,2}_{k+1/2}\varphi_{k+1} = F_k, \qquad (3.23)$$

avec

$$F_k = \int_{x_{k-1}}^{x_k} f\omega_1\,dx + \int_{x_k}^{x_{k+1}} f\omega_2\,dx = \int_{x_{k-1}}^{x_{k+1}} f\omega_k\,dx. \qquad (3.24)$$

Le problème se trouve complètement défini si l'on adjoint à (3.23) les conditions aux limites

$$\varphi_0 = 0, \quad \varphi_N = 0. \qquad (3.25)$$

On identifie (3.23) et (3.8), i.e. s'agissant des équations auto-adjointes, les méthodes de Ritz et de Galerkin (des éléments finis) conduisent aux mêmes problèmes discrétisés. La technique de Galerkin est plus universelle parce qu'elle s'applique également aux problèmes qui ne sont pas auto-adjoints.

Si les coefficients de l'équation sont des fonctions régulières, les méthodes variationnelles permettent d'obtenir des schémas aux différences d'ordre d'approximation plus élevé.

2.4. Sur la formation de sous-espaces pour les méthodes variationnelles en dimension un

Ce numéro est essentiellement destiné à ceux qui désirent pénétrer dans le détail du traitement variationnel des problèmes de la physique mathématique, ce qui explique les emprunts supplémentaires à l'analyse fonctionnelle.

On débute par une approche plus générale de la question qui permet d'obtenir des approximations aux différences de haute précision. On trouve de telles approximations dans Aubin [V], Babuška [V], Strang, Fix [III], Bramble, Schatz [V]. Nous nous bornerons à une méthode proposée par Varga [I].

On décrit ensuite un procédé de formation d'une base de fonctions trigonométriques pour la résolution variationnelle des problèmes à paramètres discontinus réguliers par morceaux.

On s'occupe enfin d'une méthode variationnelle de discrétisation par l'identité intégrale.

2.4.1. Construction de schémas variationnels aux différences de haute précision

Lorsqu'on applique les procédés variationnels à l'équation de diffusion en dimension 1

$$Lu \equiv -\frac{d}{dx} p(x) \frac{du}{dx} + r(x) \frac{du}{dx} + q(x) u = f, \qquad (4.1)$$

$$0 < x < 1,$$

avec les conditions aux limites

$$u(0) = u(1) = 0 \qquad (4.2)$$

on utilise avec succès les sous-espaces d'ordre élevé de fonctions morceaux de polynômes. Ci-dessous un schéma de construction de ces sous-espaces.

Soit donné, sur le segment $[a, b]$ de l'axe réel, un réseau régulier de pas h et à nœuds $a = x_0 < x_1 = a + h < \ldots < x_N = a + Nh < x_{N+1} = b$, avec $h = \frac{b-a}{N+1}$ et N un entier positif. On désigne par $H_N^m(a, b)$ l'ensemble des fonctions $\omega(x)$ qui sont des polynômes de degré $2m + 1$ sur chaque $[x_k, x_{k+1}] \subset [a, b]$ et vérifient pour tout $0 \leqslant i \leqslant m$ et tout $0 \leqslant k \leqslant N + 1$ les égalités

$$\omega^{(i)}(x_k) = d_{k,i},$$

$d_{k,i}$ étant des nombres donnés et $\omega^{(i)}(x)$ les dérivées i-èmes de $\omega(x)$ ($\omega^{(0)}(x) \equiv \omega(x)$).

On suppose de plus que $\omega(a) = \omega(b) = 0$, i.e. $d_{0,0} = d_{N+1,0} = 0$. Il en découle que $\omega(x) \in H_N^m(a, b)$ sont des morceaux de polynômes de $C_m(a, b)$, i.e. elles sont m fois continûment dérivables sur $[a, b]$, et $H_N^m(a, b)$ est de dimension $m(N + 2) + N$. Varga [1] a établi que cet espace a pour base le système de $m(N + 2) + N$ fonctions $\{\omega_{k,i}\}$ définies par

$$\omega_{k,0}(x_i) = \delta_{kj}, \quad \omega_{k,0}^{(i)}(x_i) = 0, \qquad (4.3)$$

$$k = 1, \ldots, N; \quad j = 0, 1, \ldots, N+1; \quad i = 1, \ldots, m;$$

$$\omega_{k,l}^{(l)}(x_i) = \delta_{kj}, \quad \omega_{k,l}^{(i)}(x_i) = 0, \qquad (4.4)$$

$$k = 0, 1, \ldots, N+1; \quad j = 0, 1, \ldots, N+1;$$

$$i = 0, 1, \ldots, l-1, l+1, \ldots, m; \quad l = 1, 2, \ldots, m,$$

δ_{kj} étant le symbole de Kronecker.

On vérifie de suite que les conditions (4.3) définissent au total N fonctions et les conditions suivantes $m(N+2)$. Comme toute $\omega(x) \in H_N^m(a, b)$ se définit par $\{d_{k,0}\}_{k=1}^N$ et $d_{k,l}$, $k = 0, \ldots, N+1$; $l = 1, \ldots, m$, on a en outre le développement

$$\omega(x) = \sum_{k=1}^{N} d_{k,0}\omega_{k,0}(x) + \sum_{k=0}^{N+1}\sum_{l=1}^{m} d_{k,l}\,\omega_{k,l}(x). \tag{4.5}$$

Voyons de plus près les fonctions $\{\omega_{k,l}\}$. Soit $m = 1$, auquel cas l'espace $H_N^1(a, b)$ est de dimension $2N+2$ et ses fonctions de base sont définies par les conditions

$$\omega_{k,0}(x_j) = \delta_{kj}, \quad \omega_{k,0}^{(1)}(x_j) = 0,$$
$$k = 1, \ldots, N; \quad j = 0, 1, \ldots, N+1; \tag{4.6}$$
$$\omega_{k,1}(x_j) = 0, \quad \omega_{k,1}^{(1)}(x_j) = \delta_{kj}, \qquad k, j = 0, 1, \ldots, N+1. \tag{4.7}$$

De la définition des fonctions $\omega_{k,l}$ il résulte qu'elles ne sont nulles que sur les intervalles $(x_k - h, x_k + h)$. La recherche de $\omega_{k,0}(x)$ sur $(x_k - h, x_k + h)$ donne lieu à deux problèmes.

Le premier problème se pose sur le segment $[x_k - h, x_k]$: on désire reconstituer le polynôme

$$\omega_{k,0}(x) = \sum_{j=0}^{3} C_j^{0,-}(x_k - x)^j \tag{4.8}$$

si $\omega_{k,0}(x_k - h) = 0$, $\omega_{k,0}^{(1)}(x_k - h) = 0$, $\omega_{k,0}(x_k) = 1$, $\omega_{k,0}^{(1)}(x_k) = 0$. On obtient le système de 4 équations pour déterminer 4 inconnues $\{C_j^{0,-}\}_{j=0}^3$:

$$\sum_{j=0}^{3} C_j^{0,-}(h)^j = 0,$$
$$\sum_{j=1}^{3} jC_j^{0,-}(h)^{j-1} = 0, \tag{4.9}$$
$$C_1^{0,-} = 0, \quad C_0^{0,-} = 1,$$

d'où

$$C_0^{0,-} = 1, \quad C_2^{0,-} = -\frac{3}{h^2},$$
$$C_1^{0,-} = 0, \quad C_3^{0,-} = \frac{2}{h^3}. \tag{4.10}$$

On procède de même pour le segment $[x_k, x_{k+1}]$ et on a

$$\omega_{k,0}(x) = \sum_{j=0}^{3} C_j^{0,+}(x_k - x)^j, \tag{4.11}$$

avec les coefficients $\{C_j^{0,+}\}_{j=0}^3$

$$C_0^{0,+}=1, \quad C_2^{0,+}=-\frac{3}{h^2}, \qquad C_1^{0,+}=0, \quad C_3^{0,+}=-\frac{2}{h^3} \tag{4.12}$$

par suite de $\omega_{k,0}(x_k+h)=0$, $\omega_{k,0}(x_k)=1$, $\omega_{k,0}^{(1)}(x_k+h)=0$, $\omega_{k,0}^{(1)}(x_k)=0$. Finalement

$$\omega_{k,0}(x)=\begin{cases} \sum\limits_{j=0}^{3} C_j^{0,-}(x_k-x)^j & \text{si} \quad x\in[x_k-h,\ x_k], \\ \sum\limits_{j=0}^{3} C_j^{0,+}(x_k-x)^j & \text{si} \quad x\in[x_k,\ x_k+h], \\ 0 \quad \text{pour les} \quad x\in[a,\ b] \quad \text{restants.} \end{cases} \tag{4.13}$$

Lorsque $a=0$, $b=2$, $N=3$, la fonction $\omega_{2,0}(x)$ est représentée par la courbe de la fig. 2.

On déduit la formule des $\omega_{k,1}(x)$ à partir de (4.7). Par des raisonnements analogues on définit cette fonction pour tout $0 \leqslant k \leqslant N+1$ par la formule

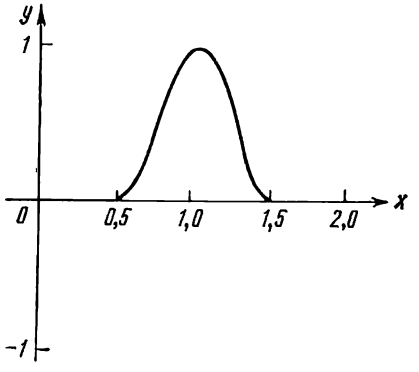


Fig. 2

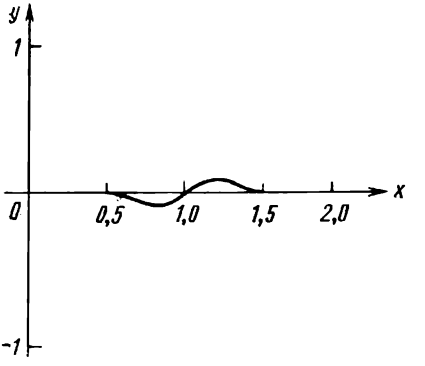


Fig. 3

$$\omega_{k,1}(x)=\begin{cases} \sum\limits_{j=0}^{3} C_j^{1,-}(x_k-x)^j & \text{si} \quad x\in[x_k-h,\ x_k]\cap[a,\ b], \\ \sum\limits_{j=0}^{3} C_j^{1,+}(x_k-x)^j & \text{si} \quad x\in[x_k,\ x_k+h]\cap[a,\ b], \\ 0 \quad \text{dans le cas contraire.} \end{cases} \tag{4.14}$$

Здесь

$$C_0^{1,-}=0, \qquad C_2^{1,-}=\frac{2}{h},$$
$$C_1^{1,-}=-1, \qquad C_3^{1,-}=-\frac{1}{h^2},$$
$$C_0^{1,+}=0, \qquad C_2^{1,+}=-\frac{2}{h}, \tag{4.15}$$
$$C_1^{1,+}=-1, \qquad C_3^{1,+}=-\frac{1}{h^2}.$$

На фиг. 3 изображен график $\omega_{2,1}(x)$ при $a=0$, $b=2$, $N=3$.

При $m=0$ формулы, дающие $\{\omega_{k,0}\}$, получаются весьма простыми. Пространство $H_N^0(a, b)$ имеет тогда в качестве элементов кусочно-линейные функции $\{\omega_k\}_{k=1}^N$, обладающие свойством

$$\omega_k(x_j)=\delta_{kj}, \qquad k, j=1, 2, \ldots, N. \tag{4.16}$$

Напомним, что $\{\omega_k(x)\}$ нам уже встречались в предыдущем параграфе (см. фиг. 1).

Приведем еще один способ построения F_h. Возьмем отрезок $[a, b]$ оси x и покроем его сеткой с узлами $a=x_0<x_1<\ldots<x_N<x_{N+1}=b$, где $x_k=a+kh$, $k=0, 1, \ldots, N+1$, $h=(b-a)/(N+1)$. Обозначим через $M_N^m(a, b)$ множество функций $g(x)$, обладающих следующими свойствами: 1) $g(x)$ на каждом $[x_k, x_{k+1}]$ является многочленом степени m; 2) для всех $0\leqslant k\leqslant N$ и $0\leqslant j\leqslant m$ выполнены равенства

$$g(x_{k,j})=d_{k,j},$$

где $x_{k,j}=x_k-\frac{h}{m}j$, а $d_{k,j}$ — заданные числа; 3) $g(a)=g(b)=0$, т. е. $d_{0,0}=d_{N,m}=0$. Отсюда следует, что $g\in M_N^m(a, b)$ принадлежит $\mathring{W}_2^1$, т. е. $g(x)$ — непрерывная функция, производные которой первого порядка могут иметь разрывы в точках $\{x_k\}_{k=1}^N$. Построим $g(x)$ явно с помощью $\{d_{k,j}\}$.

Поступаем следующим образом. Выберем произвольное $0\leqslant k\leqslant N$ и построим $g(x)$ на $[x_k, x_{k+1}]$ (обозначим результат через $g_k(x)$). Теория приближения утверждает существование и единственность многочлена степени m, проходящего через $m+1$ точек $d_{k,0}, \ldots, d_{k,m}$ плоскости x, y. Это — интерполяционный многочлен Лагранжа, вычисляемый по формуле

$$y=g_k(x)=\sum_{i=0}^{m} d_{k,i}\prod_{\substack{l=0\\ l\neq i}}^{m}\frac{x_{k,l}-x}{x_{k,l}-x_{k,i}}. \tag{4.17}$$

Si $m = 1$, on obtient une fonction linéaire usuelle, à savoir

$$g_k(x) = d_k \frac{x_{k+1} - x}{x_{k+1} - x_k} + d_{k+1} \frac{x_k - x}{x_k - x_{k+1}}. \tag{4.18}$$

Il en résulte en particulier la coïncidence de $M_N^1(a, b)$ et de $H_N^0(a, b)$, si bien qu'une base formée pour le deuxième espace en est une du premier. Nous laisserons de côté le problème de la base de $M_N^m(a, b)$ pour m importants.

Ainsi, nous avons formé pour $F = \overset{\circ}{W}{}_2^1(a, b)$ deux types de sous-espaces F_h (définis par le paramètre h) qui constituent une suite complète dans $\overset{\circ}{W}{}_2^1(a, b)$.

2.4.2. Construction d'une base de fonctions trigonométriques et son rôle dans les méthodes variationnelles

Pour nous faire une idée des principes de base de la méthode nous considérons le problème (1.1), (1.2) qui se ramène à minimiser la fonctionnelle (3.1).

Supposons que les paramètres p, q, f peuvent présenter des discontinuités de première espèce aux points $\{y_i\}_{i=1}^m$ et qu'ils sont suffisamment réguliers sur chaque $y_i \leqslant x \leqslant y_{i+1}$. On admet de plus que cette propriété des paramètres garantit sur ces intervalles la régularité jusqu'à l'ordre ν de φ, i.e. $\varphi \in Q^{(\nu)}[0, 1]$ (voir n°2.1).

Avec ν suffisamment grand, on utilise naturellement pour former la base $\{\omega_k\}_{k=1}^{\infty}$ les bonnes propriétés d'approximation des polynômes trigonométriques dont les périodes sont supérieures aux intervalles de régularité correspondants. Arrêtons-nous sur ce résultat relatif aux polynômes trigonométriques.

Une fonction $f(x) \in C^{(\nu)}[0, 1]$ définie sur $0 \leqslant x \leqslant 1$ est prolongeable (de façon non unique) aux points $x \notin [0, 1]$ en une fonction $\tilde{f}(x)$ jouissant des propriétés importantes (Fikhtengoltz [II; 11]):

$$\tilde{f}(x+T) = \tilde{f}(x), \qquad -\infty < x < \infty,$$

$\tilde{f} \in C^{(\nu)}[x_0, x_0 + T]$ quel que soit $x_0 \in (-\infty, \infty)$, $T > 1$. La vitesse de convergence de la série de Fourier de $\tilde{f}(x)$ (aussi bien dans $C[0, 1]$ que dans $L_2[0, 1]$) s'évalue par

$$\|f - \tau_N\| \leqslant \text{const} \frac{\ln N}{N^{\nu}}, \tag{4.19}$$

où

$$\tau_N = \sum_{k=0}^{N} \left(a_k \cos \frac{2\pi k x}{T} + b_k \sin \frac{2\pi k x}{T} \right).$$

On modifie les coefficients a_k et b_k (sans détériorer (4.19)) de façon à avoir en outre $\tau_N(0) = f(0)$, $\tau_N(1) = f(1)$.

A la lumière du résultat obtenu, on approche la solution de (1.1), (1.2) sur chaque intervalle de régularité $y_i \leqslant x \leqslant y_{i+1}$ par le polynôme trigonométrique

$$\varphi_{N_i}(x) \equiv \sum_{k=0}^{2N_i} C_k^i \omega_k^i(x), \qquad i = 0, 1, \ldots, m,$$

où par exemple

$$\omega_{2k-1}^i(x) = \sin \frac{k\pi(x-y_i)}{H_i}, \quad \omega_{2k}^i(x) = \cos \frac{k\pi(x-y_i)}{H_i}, \quad (4.20)$$

$$H_i = y_{i+1} - y_i.$$

Chaque $\omega_k^i(x)$ peut être prolongée par 0 à $0 \leqslant x \leqslant 1$ tout entier:

$$\omega_k^i(x) = \begin{cases} (4.20) & \text{si} \quad x \in [y_i, y_{i+1}], \\ 0 & \text{dans le cas contraire.} \end{cases} \quad (4.21)$$

On prend le système (4.21), $i = 0, \ldots, m$; $k = 0, 1, \ldots$, pour suite de fonctions de base (discontinues pour k pair) et on cherche la solution approchée du problème (3.1) sous la forme

$$\varphi(x) = \sum_{i=1}^{m} \sum_{k=0}^{2N_i} C_k^i \omega_k^i(x), \qquad 0 \leqslant x \leqslant 1. \quad (4.22)$$

Les fonctions de base (4.21) étant discontinues, leurs combinaisons linéaires (4.22) ne garantissent automatiquement pas la continuité de φ et $pd\varphi/dx$ aux points $\{y_i\}_{i=1}^m$, pas plus d'ailleurs que les conditions aux limites *). On vérifie facilement que

$$\varphi(0) = \varphi(1) = 0, \quad \varphi(x_i - 0) = \varphi(x_i + 0), \qquad i = 1, 2, \ldots, m$$

sont les *conditions essentielles* et

$$p \frac{d\varphi}{dx}\Big|_{x_i - 0} = p \frac{d\varphi}{dx}\Big|_{x_i + 0}, \qquad i = 1, 2, \ldots, m,$$

sont les *conditions naturelles* (voir Mikhlin [1]). La méthode variationnelle basée sur la minimisation de la fonctionnelle (3.1) exige qu'on soit dans les conditions essentielles, tandis que les conditions naturelles sont remplies automatiquement par $\varphi^0 \in \Phi$, qui don-

*) On vérifie de suite les conditions aux limites si l'on définit les fonctions $\omega_k^0(x)$ et $\omega_k^m(x)$ sur $[0, x_1]$ et $[x_m, 1]$ par les égalités

$$\omega_k^0(x) = \sin \frac{k\pi x}{2x_1}, \qquad \omega_k^m(x) = \sin \frac{k\pi(1-x)}{2H_m}$$

au lieu de (4.20).

ne la solution du problème variationnel. On impose donc à (4.22) de satisfaire à

$$\varphi_{N_0}(0) = \varphi_{N_m}(1) = 0, \quad \varphi_{N_{i-1}}(x_i) = \varphi_{N_i}(x_i), \qquad (4.23)$$
$$i = 1, \ldots, m.$$

On vérifie de suite l'estimation

$$\|\varphi - \varphi^0\|_{L_2[0,1]} \leqslant \text{const}\, \frac{\ln N}{N^{\nu-1}}, \qquad N = \min_i N_i \qquad (4.24)$$

pour l'écart entre la solution approchée $\varphi(x)$ et la solution exacte $\varphi^0(x)$.

Décrivons en quelques mots l'algorithme de calcul correspondant.

La minimisation de (3.1) par des fonctions de la forme (4.22) avec les conditions supplémentaires (4.23) à l'aide des multiplicateurs de Lagrange donne lieu au système

$$\left.\begin{aligned} &\sum_{k=0}^{2N_i} \alpha_k^i (A\omega_k^i, \omega_j^i) + \sum_{s=i}^{i+1} \lambda_s \beta_{js}^i = (f, \omega_j^i), \\ &i = 0, 1, \ldots, m; \quad j = 0, 1, \ldots, 2N_i, \\ &\sum_{i=s-1}^{s} \sum_{k=0}^{2N_i} \alpha_k^i \beta_{ks}^i = 0, \quad s = 0, 1, \ldots, m+1, \end{aligned}\right\} \qquad (4.25)$$

où

$$\beta_{ks}^i = \begin{cases} \omega_k^{s-1}(x_s), & i = s-1, \\ -\omega_k^s(x_s), & i = s, \ \omega_k^{-1}(x) \equiv \omega_k^{m+1}(x) \equiv 0, \\ 0, & i = s, \ s-1. \end{cases} \qquad (4.26)$$

Le système (4.25) s'écrit par blocs

$$\begin{Vmatrix} \hat{A} & B^T \\ B & 0 \end{Vmatrix} \begin{Vmatrix} X \\ \Lambda \end{Vmatrix} = \begin{Vmatrix} F \\ 0 \end{Vmatrix}, \qquad (4.27)$$

avec X formé de tous les α_k^i, Λ un vecteur dont les composantes sont des multiplicateurs de Lagrange et T l'indice de transposition.

On a à partir d (4.27)

$$\hat{A}X + B^T\Lambda = F, \quad BX = 0, \qquad (4.28)$$

d'où l'on tire formellement après élimination de X

$$(B\hat{A}^{-1}B^T)\Lambda = B\hat{A}^{-1}F. \qquad (4.29)$$

La définie positivité de $\hat{A}$ entraîne l'existence de $\hat{A}^{-1}$ et la définie positivité de $\hat{B}A^{-1}B^T$. En effet, $(B\hat{A}^{-1}B^TW, W) =$

$=(\hat{A}^{-1}B^TW, B^TW)=(\hat{A}V, V)>0$, où $V=\hat{A}^{-1}B^TW\neq 0$ pour $W\neq 0$ (l'annulation de W pour $B^TW=0$ découle de suite de (4.26) et (4.20)).

La forme bloc-diagonale de $\hat{A}$ donne sans peine

$$B\hat{A}^{-1}B^T=\sum_{i=0}^{m} B_iA_i^{-1}B_i^T, \qquad BA^{-1}F=\sum_{i=0}^{m} B_iA_i^{-1}F_i, \tag{4.30}$$

A_i et B_i étant deux matrices d'éléments respectifs

$$(A\omega_k^i,\ \omega_j^i) \text{ et } (\beta_{ks}^i)$$

et F_i des vecteurs de composantes

$$(f,\ \omega_0^i),\ \ldots,\ (f,\ \omega_{2N_i}^i).$$

Selon (4.30), l'inversion de $\hat{A}$ se ramène à celle de tous ses blocs, autrement dit, on a en général à inverser des matrices d'ordre peu élevé. Si m est petit, il en est également de l'ordre de (4.29), si bien que ce système à matrice symétrique définie positive est résolu sans difficulté.

On trouve enfin moyennant la première équation (4.28)

$$X_i=A_i^{-1}(F_i-B_i^T\Lambda), \tag{4.31}$$

où

$$X_i=(\alpha_0^i,\ \ldots,\ \alpha_{2N_i}^i)^T, \qquad i=0,\ 1,\ \ldots,\ m.$$

2.4.3. Identité intégrale sous forme variationnelle

Nous avons examiné plus haut (n°2.1) un procédé de construction d'équations aux différences qui utilise une identité intégrale. Agochkov [V; 1] en a montré le caractère variationnel. Soit en effet l'équation différentielle

$$-\frac{d}{dx}p(x)\frac{du}{dx}+q(x)u=f(x), \qquad x\in(a,\ b), \tag{4.32}$$

avec les conditions aux limites

$$u(a)=u(b)=0. \tag{4.33}$$

On suppose que $p(x)>0$, $q(x)\geqslant 0$, $p(x)$, $q(x)\in L_\infty(a,\ b)$, $f(x)\in L_2(a,\ b)$. S'agissant du problème (4.32), (4.33), la méthode

mentionnée consiste à obtenir, compte tenu de (4.32), (4.33), l'identité

$$\frac{u(x_k)-u(x_{k+1})}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k)-u(x_{k-1})}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} + \int\limits_{x_{k-1/2}}^{x_{k+1/2}} (qu-f)\,dx =$$

$$= -\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)} \int\limits_{x_{k+1/2}}^{x} (qu-f)\,d\xi +$$

$$+ \frac{1}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)} \int\limits_{x_{k-1/2}}^{x} (qu-f)\,d\xi, \quad (4.34)$$

avec $a = x_0 < x_{1/2} < x_1 < x_{3/2} < \ldots < x_{N-3/2} < x_{N-1} < x_{N-1/2} < x_N = b$ un ensemble de points. Cherchons une approximation quelconque des intégrales de (4.34), il vient le schéma aux différences correspondant.

Notre but immédiat est de montrer que (4.34) peut se mettre sous une forme variationnelle proche des équations variationnelles de la méthode de Galerkin à la différence qu'elle permet d'utiliser des fonctions de base discontinues et de démontrer la convergence des solutions approchées sous des hypothèses assez générales sur la régularité des données initiales.

Effectuons dans (4.34) des transformations élémentaires. Soit

$\psi = qu - f$, $\rho_k(x) = \int\limits_{x_k}^{x} \frac{dx'}{p(x')} \Big/ \int\limits_{x_k}^{x_{k+1}} \frac{dx'}{p(x')}$. Alors

$$-\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)} \int\limits_{x_{k+1/2}}^{x} \psi(\xi)\,d\xi =$$

$$= -\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int\limits_{x_k}^{x_{k+1}} d\left(\int\limits_{x_k}^{x} \frac{dx'}{p(x')}\right) \int\limits_{x_{k+1/2}}^{x} \psi(\xi)\,d\xi =$$

$$= -\int_{x_{k+1/2}}^{x_{k+1}} \psi(\xi)\, d\xi + \frac{1}{\int_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int_{x_k}^{x_{k+1}} \psi(x) \int_{x_k}^{x} \frac{d\xi}{p(\xi)}\, dx =$$

$$= -\int_{x_{k+1/2}}^{x_{k+1}} \psi(\xi)\, d\xi + \int_{x_k}^{x_{k+1}} \rho_k(x)\, \psi(x)\, dx.$$

On obtient de même

$$\frac{1}{\int_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} \int_{x_{k-1}}^{x_k} \frac{dx}{p(x)} \int_{x_{k-1/2}}^{x} \psi(\xi)\, d\xi =$$

$$= -\int_{x_{k-1}}^{x_{k-1/2}} \psi(x)\, dx + \int_{x_{k-1}}^{x_k} \tilde{\rho}_k(x)\, \psi(x)\, dx,$$

où $\tilde{\rho}_k(x) = \int_x^{x_k} \frac{dx'}{p(x')} \Big/ \int_{x_{k-1}}^{x_k} \frac{dx}{p(x)}$. L'identité (4.34) se récrit

$$\frac{u(x_k) - u(x_{k+1})}{\int_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k) - u(x_{k-1})}{\int_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} +$$

$$+ \int_{x_{k-1}}^{x_k} (1 - \tilde{\rho}_k(x))\, \psi(x)\, dx + \int_{x_k}^{x_{k+1}} (1 - \rho_k(x))\, \psi(x)\, dx = 0.$$

Soit

$$Q_k(x) = \begin{cases} 1 - \int_x^{x_k} \frac{d\xi}{p(\xi)} \Big/ \int_{x_{k-1}}^{x_k} \frac{d\xi}{p(\xi)}, & x \in (x_{k-1},\ x_k), \\ 1 - \int_{x_k}^{x} \frac{d\xi}{p(\xi)} \Big/ \int_{x_k}^{x_{k+1}} \frac{d\xi}{p(\xi)}, & x \in (x_k,\ x_{k+1}), \\ 0, & x \notin (x_{k-1},\ x_{k+1}), \end{cases} \tag{4.35}$$

auquel cas (4.34) prend la forme

$$\frac{u(x_k)-u(x_{k+1})}{\int_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k)-u(x_{k-1})}{\int_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} + (qu, Q_k) = (f, Q_k), \qquad (4.36)$$

$$k = 1, \ldots, N-1,$$

avec $(\varphi, \psi) = \int_a^b \varphi \cdot \psi \, dx$, $\|\varphi\| = (\varphi, \varphi)^{1/2}$. On observe que

$$\left(p \frac{du}{dx}, \frac{dQ_k}{dx}\right) = \frac{u(x_k)-u(x_{k+1})}{\int_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k)-u(x_{k-1})}{\int_{x_{k-1}}^{x_k} \frac{dx}{p(x)}},$$

si bien qu'on a, au lieu de (4.36),

$$\left(p \frac{du}{dx}, \frac{dQ_k}{dx}\right) + (qu, Q_k) = (f, Q_k), \qquad k = 1, \ldots, N-1. \qquad (4.37)$$

Introduisons un interpolant $u_I(x)$ de $u(x)$ tel que $u_I(x_i) = u(x_i)$, $i = 0, \ldots, N$, et la dérivée du_I/dx ait un sens. En vertu de l'égalité $\left(p \frac{du}{dx}, \frac{dQ_k}{dx}\right) = \left(p \frac{du_I}{dx}, \frac{dQ_k}{dx}\right)$, l'expression (4.37) est équivalente à

$$\left(p \frac{du_I}{dx}, \frac{dQ_k}{dx}\right) + (qu, Q_k) = (f, Q_k), \qquad k = 1, \ldots, N-1. \qquad (4.38)$$

Ainsi, l'identité (4.34) est équivalente à chacune des équations (4.36) à (4.38), ce qui autorise à ranger la méthode de l'identité intégrale parmi les techniques variationnelles et à utiliser (4.36) à (4.38) et (4.34) pour construire les solutions approchées sous des hypothèses assez lâches sur les fonctions de base. La relation (4.37) n'est rien d'autre que l'égalité variationnelle connue que la méthode de Galerkin emploie en cas de fonctions de base $Q_k(x)$. S'agissant de (4.36), il y a lieu de dire qu'on peut utiliser des fonctions de base avec discontinuités (qui ne coïncident pas avec x_i, $i = 0, \ldots, N$). Avec l'identité (4.38), on obtient sans trop de peine plusieurs estimations dans la métrique uniforme.

Essayons d'approcher le problème (4.32), (4.33) moyennant les identités (4.36) à (4.38) et diverses fonctions de base. Le changement de variable $y = \int_a^x \frac{d\xi}{p(\xi)}$ détermine l'écriture suivante du pro-

blème proposé:

$$\begin{cases} -\dfrac{d^2u}{dy^2}+p(y)\,q(y)\,u=\widetilde{f}(y), & 0<y<T=\displaystyle\int_a^b \frac{d\xi}{p(\xi)}, \\ u(0)=u(T)=0. \end{cases}$$

Si $f(x)\in L_2(a, b)$, on a donc $u(y)\in W_{2,0}^2(0, T)$. On cherche maintenant une solution approchée de la forme $u^h(x)=\sum\limits_{i=1}^{N-1} a_iQ_i(x)$, les inconnues a_i étant définies à partir du système

$$\left(p\frac{du^h}{dx},\ \frac{dQ_k}{dx}\right)+(qu^h,\ Q_k)=(f,\ Q_k),\qquad k=1,\ \ldots,\ N-1,$$

qui admet une solution unique. Autrement dit, on emploie la tactique usuelle de Galerkin avec les fonctions de base $\{Q_i(x)\}$. Comme $Q_i(x)$ deviennent par la substitution $y=\int\limits_a^x \frac{d\xi}{p(\xi)}$ des « chapeaux pointus » multi-linéaires ordinaires, on peut dire, compte tenu des résultats connus de la théorie de l'approximation, que la solution approchée tend pour $h=\max\limits_i |x_i-x_{i-1}|\to 0$ vers la solution exacte et qu'on a les estimations

$$\|u-u^h\|_{W_2^k}\leqslant O(h^{2-k}),\qquad k=0,\ 1.$$

Mais la relation (4.38) avec $u_{\mathrm{I}}=\sum\limits_{i=1}^{N-1} u(x_i)\,Q_i(x)$ est, elle aussi, valide pour la solution exacte, si bien que

$$\left(p\frac{d(u_{\mathrm{I}}-u^h)}{dx},\ \frac{d(u_{\mathrm{I}}-u^h)}{dx}\right)+(q(u-u^h),\ u_{\mathrm{I}}-u^h)=0,$$

$$\left(p\frac{d(u_{\mathrm{I}}-u^h)}{dx},\ \frac{d(u_{\mathrm{I}}-u^h)}{dx}\right)+(q(u-u^h),\ u-u^h)=$$
$$=(q(u-u^h),\ u-u_{\mathrm{I}})\leqslant(q(u-u^h),\ u-u^h)^{1/2}\times$$
$$\times(q(u-u_{\mathrm{I}}),\ u-u_{\mathrm{I}})^{1/2}.$$

Etant donné $\|u-u_{\mathrm{I}}\|\leqslant Ch^2$, on en tire

$$\left(p\frac{d(u_{\mathrm{I}}-u^h)}{dx},\ \frac{d(u_{\mathrm{I}}-u^h)}{dx}\right)+(q(u-u^h),\ u-u^h)\leqslant Ch^4.$$

On a de plus

$$\|u-u^h\|_{C(a,\,b)}\leqslant Ch^2$$

du moment que

$$\max_{x\in(x_{i-1},\ x_i)}|u(x)-u_{\mathrm{I}}(x)|\leqslant Ch^2\left\|\frac{d^2u}{dy^2}\right\|_{L_\infty(y_{i-1},\ y_i)}\leqslant Ch^2,$$

pour $f\in L_\infty$.

Considérons deux cas avec les fonctions de base $\{\varphi_i^h(x)\}$.

Soit $h_i = x_{i+1/2} - x_{i-1/2}$, $h = \max\limits_i h_i$. Désignons par $\varphi_i^h(x)$ la fonction caractéristique de l'intervalle $(x_{i-1/2},\ x_{i+1/2})$, par $\varphi_0^h(x)$ celle de $(x_0,\ x_{1/2})$ et par $\varphi_N^h(x)$ celle de $(x_{N-1/2},\ x_N)$. Prenons $\{\varphi_i^h\}_{i=0}^N$ pour fonctions de base et cherchons la solution approchée sous la forme $u^h(x) = \sum\limits_{i=0}^{N} a_i \varphi_i^h(x)$ en posant $a_0 = a_N = 0$. Alors $u^h(x) = \sum\limits_{i=1}^{N-1} a_i \varphi_i^h(x)$, où $\{a_i\}_{i=1}^{N-1}$ sont définis moyennant (4.36) :

$$\frac{u^h(x_k) - u^h(x_{k+1})}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u^h(x_k) - u^h(x_{k-1})}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} + (qu^h,\ Q_k) = (f,\ Q_k), \quad (4.39)$$

$$k = 1, \ldots, N-1.$$

Les calculs simples prouvent qu'il y a unicité pour (4.39) avec h suffisamment petit, que les solutions approchées u^h convergent vers $u(x)$ quand $h \to 0$ et qu'on a les estimations

$$\left(p\,\frac{d(u_I - u_I^h}{dx},\ \frac{d(u_I - u_I^h)}{dx}\right) + (q(u - u^h),\ u - u^h) \leqslant Ch,$$

$$\max_i |u(x_i) - u^h(x_i)| \leqslant Ch,$$

où

$$u_I(x) = \sum_{i=1}^{N-1} u(x_i)\, Q_i(x), \qquad u_I^h(x) = \sum_{i=1}^{N-1} a_i Q_i(x).$$

On utilise également les chapeaux pointus $\{\varphi_i^h(x)\}_{i=1}^{N-1}$ linéaires par morceaux sur $[a,\ b]$, $\varphi_i^h(x)$ étant nulle en dehors de $(x_{i-1},\ x_{i+1})$ et $\varphi_i^h(x_i) = 1$. On cherche une solution de la forme $u^h(x) = \sum\limits_{i=1}^{N-1} a_i \varphi_i^h(x)$, les inconnues a_i étant définies par le système d'équations algébriques linéaires

$$\left(p\,\frac{du^h}{dx},\ \frac{dQ_i}{dx}\right) + (qu^h,\ Q_i) = (f,\ Q_i), \qquad i = 1, \ldots, N-1. \quad (4.40)$$

Sous l'hypothèse de $p(x)$ régulière par morceaux avec des discontinuités éventuelles de première espèce en certains nœuds x_i, on

montre que le système (4.40) possède pour $h = \max_i |x_i - x_{i-1}|$ suffisamment petits une solution unique $\{a_i\}_{i=1}^{N-1}$ et qu'on a l'estimation d'erreur

$$\left(p\frac{d(u_I - u_I^h)}{dx}, \frac{d(u_I - u_I^h)}{dx}\right) + (q(u-u^h),\ u-u^h) \leqslant Ch^4 \|f\|^2;$$

si $f(x) \in L_\infty(a, b)$, on démontre tout aussi facilement

$$\|u - u^h\|_{C(a,b)} \leqslant Ch^2 \|f\|_{L_\infty}.$$

Ce procédé d'approximation avec l'identité intégrale sous forme variationnelle s'applique (le fait est prouvé par Agochkov [V; 1]) à plusieurs autres problèmes de la physique mathématique.

2.5. Schémas variationnels aux différences pour une équation elliptique en dimension deux

2.5.1. Méthode de Ritz

Nous allons construire des schémas variationnels aux différences pour le problème (2.1), (2.2) sous les hypothèses supplémentaires :

$$B_i(x) \equiv 0 \quad \text{dans} \quad D, \qquad i = 1,\ 2;$$

$$q(x) \equiv 0 \ \text{dans}\ D;$$

tout vecteur $\xi = (\xi_1, \xi_2)$ vérifie l'inégalité

$$\mu_0 \sum_{i=1}^{2} \xi_i^2 \leqslant \inf_{x \in D} \sum_{i,j=1}^{2} A_{ij}(x)\xi_i\xi_j \leqslant \sup_{x \in D} \sum_{i,j=1}^{2} A_{ij}(x)\xi_i\xi_j \leqslant \mu_1 \sum_{i=1}^{2} \xi_i^2 \tag{5.1}$$

avec des constantes positives $\mu_0 \leqslant \mu_1$; la frontière ∂D du domaine D est linéaire par morceaux. Il faut dire que la méthode de Ritz joue avec les trois dernières contraintes affaiblies, et si nous ne le faisons pas, c'est uniquement pour simplifier les calculs. Ainsi, on a

$$-\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} A_{ij}(x) \frac{\partial u}{dx_j} = f \quad \text{dans} \quad D \tag{5.2}$$

avec la condition aux limites

$$u = 0 \ \text{sur}\ \partial D. \tag{5.3}$$

Ce problème équivaut, selon le n°2.2, à chercher la fonction qui minimise dans $\overset{\circ}{W}{}_2^1(D)$ la fonctionnelle quadratique

$$J(u) = \int_D \left[\sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j}\right] dD - 2\int_D uf\, dD. \tag{5.4}$$

On approche le problème proposé par la méthode de Ritz avec des sous-espaces F_h spéciaux. On construit F_h en triangulant D, i.e. on recouvre D par un réseau D_h à mailles triangulaires (fig. 4). On construit dans chaque triangle ainsi obtenu un polynôme de degré m en x_1, x_2 tel que

$$g(x_1, x_2) = \sum_{i=0}^{m} \sum_{j_1+j_2=i} c_{j_1 j_2} x_1^{j_1} x_2^{j_2}. \tag{5.5}$$

Dans chaque triangle les coefficients du polynôme sont choisis de façon que toute la fonction soit dans $\overset{\circ}{W}{}_2^1(D)$, i.e. elle doit être continue et s'annuler sur ∂D.

Pour $m = 1$, la méthode a été proposée par Courant [V]. Elle a été étudiée par Oganessian [V; 7] et d'autres auteurs. Une analyse approfondie des cas $m = 2, 3, 5$ est due à Zlamal [V].

Nous nous intéresserons au cas $m = 1$, i.e. à la fonction linéaire par morceaux

$$g(x_1, x_2) = c_{0,0} + c_{1,0} x_1 + c_{0,1} x_2. \tag{5.6}$$

On détermine ses coefficients pour un triangle concret par les valeurs données $u(p_1)$, $u(p_2)$, $u(p_3)$ aux sommets p_1, p_2, p_3. On répète l'opération autant de fois que D contient de triangles en posant $u(p) = 0$ pour les points $p \in \partial D$ et on constate la continuité et la

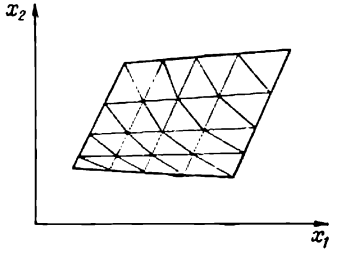


Fig. 4

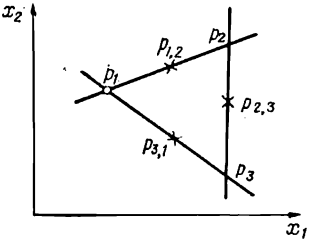


Fig. 5

nullité sur ∂D de la résultante. La continuité de cette fonction sur les côtés communs des triangles (par exemple sur le segment joignant p_1 à p_2) tient à ce que $g(x_1, x_2)$ sur ces côtés constitue chaque fois une droite d'extrémités $u(p_1)$ et $u(p_2)$.

Si l'on se place dans le cas de l'interpolation quadratique ($m = 2$)

$$g(x_1, x_2) = c_{0,0} + c_{1,0} x_2 + c_{0,1} x_2 + c_{2,0} x_1^2 + c_{1,1} x_1 x_2 + c_{0,2} x_2^2, \tag{5.7}$$

il y a intérêt à garantir l'appartenance $u^h(x_1, x_2) \in \overset{\circ}{W}{}_2^1$. On se donne à cet effet les valeurs de u^h aux sommets p_1, p_2, p_3 et aux points $p_{1,2}$, $p_{2,3}$, $p_{3,1}$, milieux des segments (p_1, p_2), (p_2, p_3), (p_3, p_1) (fig. 5).

La continuité de u^h tient à ce que cette fonction est définie univoquement sur chaque côté de tout triangle. Par exemple $u^h(x_1, x_2)$ est définie sur $[p_1, p_2]$ par les valeurs $u(p_1)$, $u(p_{1,2})$, $u(p_2)$, et ce de façon unique.

Si l'on cherche la solution dans chaque domaine triangulaire Δ par la représentation polynomiale (5.7), on est conduit par la méthode de Ritz classique et par recours à une fonctionnelle variationnelle à un système d'équations aux différences. Zlamal a démontré que si la solution $\hat{u}$ du problème (5.2), (5.3) est de classe $C^{(3)}(D)$, les dérivées troisièmes de $\hat{u}(x)$ sont bornées en valeur absolue par une quantité M et le plus petit angle des triangles constitutifs est minoré par $\nu_0 > 0$, alors l'erreur sur la solution approchée selon Ritz du problème (5.2), (5.3) est évaluée dans le cas (5.7) par l'inégalité

$$\| u^h - \hat{u} \|_{\mathring{W}_2^1} \leqslant Ch^2,$$

où $C = \frac{C_1 M}{\sin \nu_0}$ et la constante C_1 ne dépend pas de la triangulation de D.

Si la solution du problème (5.2), (5.3) est assujettie à être moins régulière ($\hat{u} \in C^{(2)}(D)$), on démontre facilement la majoration suivante pour les approximations linéaires par morceaux :

$$\| u^h - \hat{u} \|_{\mathring{W}_2^1} \leqslant Ch,$$

$C > 0$ étant une constante.

Il nous paraît inutile de nous attarder sur les approximations quadratiques dans les domaines triangulaires (nous en avons dit assez pour pouvoir déduire les systèmes algébriques concrets). Les approximations linéaires par morceaux méritent par contre une analyse plus attentive. Les résultats antérieurs entraînent qu'une fonction $g(x) \in \mathring{W}_2^1(D)$ linéaire par morceaux dans chaque triangle est définie de façon unique par ses valeurs aux sommets. Désignons par $\{p_k\}_{k=1}^{Nh}$ l'ensemble des sommets intérieurs à D et par $D_{h,k}$ la réunion des triangles de sommet en p_k. L'espace F_h a pour base un système de fonctions $\{\omega_k(x)\}_{k=1}^{Nh}$ satisfaisant aux conditions :

1) $\omega_k(p_j) = \delta_{kj}$, δ_{kj} étant le symbole de Kronecker ;

2) $\omega_k(x)$ est linéaire dans chaque triangle, i.e. on a la représentation (5.6).

Ainsi, la fonction $\omega_k(x)$ s'interprète géométriquement comme une pyramide de sommet p_k et elle s'annule partout sauf à l'intérieur de $D_{h,k}$ (fig. 6).

Soit, en particulier, $D = \{(x_1, x_2);\ 0 < x_1, x_2 < 1\}$ le carré unité. On le recouvre d'un réseau régulier usuel à mailles carrées de

longueur $h = \frac{1}{N+1}$ (N est un entier positif) et on introduit sur D la triangulation de la fig. 7.

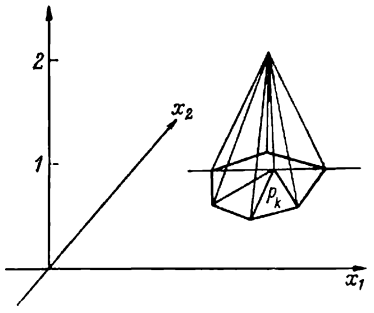


Fig. 6

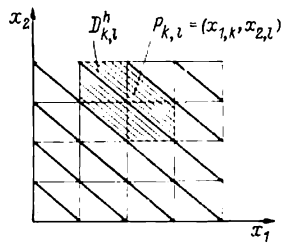


Fig. 7

Le système de fonctions de base $\{\omega_k(x)\}_{k=1}^{N^2}$, $N_h = N^2$, est noté en l'occurrence $\{\omega_{k,l}(x)\}_{k,\,l=1}^{N}$.

Etablissons certaines propriétés de la matrice A du système linéaire

$$A\alpha = g \tag{5.8}$$

de la méthode de Ritz. Ici $\alpha = (\alpha_1, \ldots, \alpha_{N^2})'$ est un vecteur formé des coefficients $\{\alpha_{N(k-1)+l} = \alpha_{k,\,l}\}_{k,\,l=1}^{N}$ du développement

$$u^h(x) = \sum_{k,\,l=1}^{N} \alpha_{k,\,l}\omega_{k,\,l}(x), \tag{5.9}$$

$g = (g_1, \ldots, g_{N^2})'$ un vecteur de composantes

$$g_{N(k-1)+l} = g_{k,\,l} = \int\limits_{D_{k,\,l}} f\omega_{k,\,l}(x)\,dD, \qquad k,\ l = 1, \ldots, N, \tag{5.10}$$

les éléments de A étant calculés par les formules

$$a_{N(k-1)+l,\,N(i-1)+j} = \int\limits_{D} \sum_{s,\,t=1}^{2} A_{s,\,t}(x)\,\frac{\partial\omega_{k,\,l}}{\partial x_s}\,\frac{\partial\omega_{i,\,j}}{\partial x_t}\,dD, \tag{5.11}$$

$$k,\ l,\ i,\ j = 1, \ldots, N.$$

Introduisons la notation $a_{k,\,l}^{i,\,j} = a_{N(k-1)+l,\,N(i-1)+j}$.

L'allure des fonctions $\{\omega_{k,\,l}(x)\}_{k,\,l=1}^{N}$ (voir fig. 7) aidant, on montre sans peine que

$$a_{k,\,l}^{i,\,j} = 0$$

si l'une au moins des inégalités

$$|i-k|>1, \qquad |j-l|>1, \qquad k, l, i, j = 1, \ldots, N,$$

a lieu. Il en résulte de suite que A est une matrice tridiagonale par blocs de la forme

$$A = \left\| \begin{array}{ccccccc} A_{11} & A_{12} & 0 & & 0 & 0 \\ A_{21} & A_{22} & A_{23} & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & A_{N,N-1} & A_{NN} \end{array} \right\|, \tag{5.12}$$

où $A_{kk}=A_{kk}^*$, $A_{k,k+1}=A_{k+1,k}^*$, $k=1, \ldots, N$, chaque $A_{k,l}$ étant tridiagonale d'ordre N. Une analyse plus scrupuleuse révèle que les matrices $\{A_{k,k-1}\}_{k=2}^{N}$ sont bidiagonales de la forme

$$A_{k,k-1} = \left\| \begin{array}{ccccccc} a_{k,1}^{k-1,1} & a_{k,1}^{k-1,2} & 0 & \ldots & 0 & 0 \\ 0 & a_{k,2}^{k-1,2} & a_{k,2}^{k-1,3} & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & a_{k,N}^{k-1,N} \end{array} \right\|, \tag{5.13}$$

$$k = 2, \ldots, N.$$

Calculons les éléments $\{a_{k,l}^{i,j}\}$ de A pour le cas particulier (5.2), (5.3) :

$$-\frac{\partial}{\partial x} p(x, y) \frac{\partial u}{\partial x} - \frac{\partial}{\partial y} q(x, y) \frac{\partial u}{\partial y} = f \text{ dans } D, \tag{5.14}$$

$$u = 0 \quad \text{sur} \quad \partial D.$$

A cet effet, on écrit $D_{k,l}^h$ comme réunion de six triangles $\{D_{k,l,m}^h\}_{m=1}^6$ numérotés comme l'indique la fig. 8. Les calculs directs donnent $\omega_{k,l}(x, y) =$

$$= \begin{cases} 1-\frac{1}{h}(x_k-x)-\frac{1}{h}(y_l-y), & x, y \in D_{k,l,1}^h, \\ 1-\frac{1}{h}(x_k-x), & x, y \in D_{k,l,2}^h, \\ 1+\frac{1}{h}(y_l-y), & x, y \in D_{k,l,3}^h, \\ 1+\frac{1}{h}(x_k-x)+\frac{1}{h}(y_l-y), & x, y \in D_{k,l,4}^h, \\ 1+\frac{1}{h}(x_k-x), & x, y \in D_{k,l,5}^h, \\ 1-\frac{1}{h}(y_l-y), & x, y \in D_{k,l,6}^h. \end{cases} \tag{5.15}$$

Etant données la symétrie de A, la propriété de triple-diagonalité des $\{A_{k,k}\}_{k=1}^{N}$ et la bi-diagonalité des $\{A_{k,k-1}\}_{k=2}^{N}$ et des

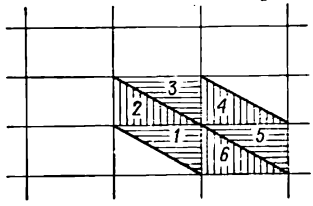


Fig. 8

$\{A_{k,k+1}\}_{k=1}^{N-1}$, il suffit de déduire les formules pour

$$a_{k,l}^{k,l},\quad a_{k,l}^{k,l-1},\quad a_{k,l}^{k-1,l},\quad a_{k,l}^{k-1,l-1},\qquad 1\leqslant k,\ l\leqslant N.$$

Selon (5.11) et (5.15), ces formules s'écrivent (pour plus de simplicité, $D_i = D_{k,l,i}^{h}$)

$$a_{k,l}^{k,l} = \int\limits_{D_{k,l}^{h}} \left[p(x,y)\left(\frac{\partial\omega_{k,l}}{\partial x}\right)^2 + q(x,y)\left(\frac{\partial\omega_{k,l}}{\partial y}\right)^2\right] dx\,dy =$$

$$= \frac{1}{h^2}\left[\int\limits_{D_1\cup D_2} p(x,y)\,dx\,dy + \int\limits_{D_4\cup D_5} p(x,y)\,dx\,dy + \right.$$

$$\left. + \int\limits_{D_3\cup D_4} q(x,y)\,dx\,dy + \int\limits_{D_1\cup D_6} q(x,y)\,dx\,dy\right], \quad (5.16)$$

$$a_{k,l}^{k,l-1} = \int\limits_{D_{k,l}^{h}\cup D_{k,l-1}^{h}} \left[p(x,y)\frac{\partial\omega_{k,l}}{\partial x}\frac{\partial\omega_{k,l-1}}{\partial x} + q(x,y)\frac{\partial\omega_{k,l}}{\partial y}\times\right.$$

$$\left.\times\frac{\partial\omega_{k,l-1}}{\partial y}\right] dx\,dy = -\frac{1}{h^2}\left[\int\limits_{D_1\cup D_6} q(x,y)\,dx\,dy\right],$$

$$a_{k,l}^{k-1,l} = -\frac{1}{h^2}\left[\int\limits_{D_1\cup D_2} p(x,y)\,dx\,dy\right], \qquad a_{k,l}^{k-1,l+1} = 0.$$

Il en résulte de suite la diagonalité des matrices $\{A_{k,k-1}\}_{k=1}^{N}$. Si l'on introduit de plus le vecteur u de composantes $u_{k,l} = \alpha_{k,l}$ avec α le vecteur du système (5.8) de la méthode de Ritz, le système $Au = g$ prend la forme

$$(A_1 + A_2)\,u = g, \quad (5.17)$$

où

$$(A_1u)_{k,\,l} = \\ = -\widetilde{P}_{k-1/2,\,l}u_{k-1,\,l} + (\widetilde{P}_{k-1/2,\,l} + \widetilde{P}_{k+1/2,\,l})\,u_{k,\,l} - \widetilde{P}_{k+1/2,\,l}u_{k+1,\,l}$$

et (5.18)

$$(A_2u)_{k,\,l} = -\widetilde{Q}_{k,\,l-1/2}u_{k,\,l-1} + (\widetilde{Q}_{k,\,l-1/2} + \widetilde{Q}_{k,\,l+1/2})\,u_{k,\,l} - \\ - \widetilde{Q}_{k,\,l+1/2}u_{k,\,l+1}.$$

Ici

$$\widetilde{P}_{k\pm1/2,\,l} = \frac{1}{h^2} \int\limits_{D^h_{k,\,l}\cap D^h_{k\pm1},} p\,(x,\,y)\,dx\,dy,$$
$$\widetilde{Q}_{k,\,l\pm1/2} = \frac{1}{h^2} \int\limits_{D^h_{k,\,l}\cap D^h_{k,\,l\pm1}} q\,(x,\,y)\,dx\,dy. \tag{5.19}$$

On réunit (5.17), (5.18), (5.19) et on vérifie sans peine que le schéma variationnel aux différences obtenu par la technique de Ritz coïncide en fait, pour ce qui est de la disposition et de la forme des éléments non nuls, avec les schémas purement discrétisés. En particulier, il n'y a aucune différence, pour $p\,(x,\,y)$ et $q\,(x,\,y)$ constants, entre l'analogue discrétisé et l'analogue variationnel aux différences de l'opérateur différentiel. Cette circonstance autorise à recourir, pour résoudre le système (5.17), à des méthodes itératives efficaces telles que la décomposition et la surrelaxation successive.

2.5.2. Méthode de Galerkin

Nous avons décrit pour la méthode de Ritz les particularités essentielles des algorithmes de résolution en bi-dimensionnel, si bien qu'il n'est plus nécessaire d'étudier en détail la formation de schémas variationnels aux différences dans le cas de Galerkin. Il suffit de dire que si le problème (2.1), (2.2) a ses coefficients $\{B_i\,(x)\}_{i=1}^2$ non nuls, la matrice $\widetilde{A}$ de la méthode de Galerkin avec les fonctions de base (5.15) diffère de la matrice A du système (5.17) par une matrice B: $B = \widetilde{A} - A$. Les éléments $\{b^{i,\,j}_{k,\,l}\}$ de B sont définis par la formule

$$b^{i,\,j}_{k,\,l} = \int\limits_{D^h_{k,\,l}\cap D^h_{i,\,j}} \left[B_1\,(x,\,y)\,\frac{\partial\omega_{ij}}{\partial x}\,\omega_{k,\,l} + B_2\,(x,\,y)\,\frac{\partial\omega_{ij}}{\partial y}\,\omega_{k,\,l}\right] dx\,dy, \tag{5.20}$$

$$k,\ l,\ i,\ j = 1,\ \ldots,\ N.$$

Conclusion : même pour l'équation

$$\sum_{i=1}^{2}\left[-\frac{\partial^2 u}{\partial x_i^2}+B_i(x)\,u\right]=f \quad \text{dans} \quad D, \tag{5.21}$$

$$u=0 \quad \text{sur } \partial D,$$

les matrices $\{A_{k,\,k-1}\}_{k=2}^{N}$ de la méthode de Galerkin peuvent ne pas être diagonales.

La forme plus complexe de la matrice du système algébrique exige des sous-espaces F_h plus simples que ceux de la méthode de Ritz. La nécessité de ces F_h se fera également sentir lorsque nous nous proposerons de construire des schémas variationnels aux différences d'ordre de précision supérieur à un.

S'agissant des domaines qui constituent les réunions de rectangles en nombre fini, on construit aisément lesdits espaces à la lumière des résultats du n° 2.4. Nous allons décrire la structure des F_h pour des approximations linéaires par morceaux et former un schéma variationnel aux différences pour un cas particulier.

Soit D la réunion de r rectangles $\{D_i\}_{i=1}^{r}$ de côtés parallèles aux axes de coordonnées et $\widetilde{D}$ le plus petit rectangle contenant D (fig. 9),

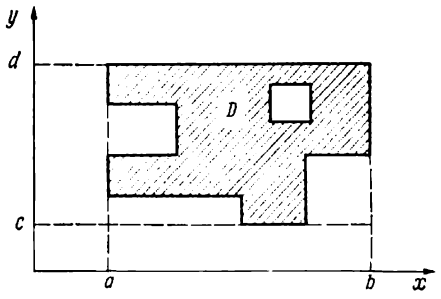


Fig. 9

où $\widetilde{D}=\{(x,\ y);\ a\leqslant x\leqslant b,\ c\leqslant y\leqslant d\}$. Superposons aux segments $[a,\ b]$ et $[c,\ d]$ deux réseaux respectifs $a=x_0<x_1<\ldots$ $\ldots<x_{N+1}=b$ et $c=y_0<y_1<\ldots<y_{M+1}=d$ tels que les rectangles de D aient nécessairement leurs côtés confondus avec une des droites

$$x=x_k,\ y=y_l,\ k=0,\ 1,\ \ldots,N+1;\ l=0,\ 1,\ \ldots,M+1. \tag{5.22}$$

On définit D_h comme étant l'ensemble des points $(x_k,\ y_l)\in D$, $k=1,\ \ldots,\ N;\ l=1,\ \ldots,\ M$. Construisons $F_h\subset \overset{\circ}{W}{}_2^1(D)$.

On introduit les fonctions

$$\omega_{x,\,k}(x)=\begin{cases}\dfrac{x-x_{k-1}}{x_k-x_{k-1}} & \text{si} \quad x\in[x_{k-1},\ x_k],\\ \dfrac{x-x_{k+1}}{x_k-x_{k+1}} & \text{si} \quad x\in[x_k,\ x_{k+1}],\\ 0 \text{ partout ailleurs,}\end{cases} \tag{5.23}$$

$$k=1,\ \ldots,\ N,$$

$$\omega_{y,\,l}(y)=\begin{cases}\dfrac{y-y_{l-1}}{y_l-y_{l-1}} & \text{si} \quad y\in[y_{l-1},\ y_l],\\ \dfrac{y-y_{l+1}}{y_l-y_{l+1}} & \text{si} \quad y\in[y_l,\ y_{l+1}],\\ 0 \text{ partout ailleurs,}\end{cases} \tag{5.24}$$

$$l=1,\ \ldots,\ M,$$

et le système de fonctions

$$\omega_{k,\,l}(x,\ y)=\omega_{x,\,k}(x)\,\omega_{y,\,l}(y), \qquad (x_k,\ y_l)\in D_h \tag{5.25}$$

et on prend pour F_h l'enveloppe linéaire des $\{\omega_{k,l}\}$. Le système $\{\omega_{k,l}\}$ étant linéairement indépendant constitue évidemment une base de F_h.

Ecrivons le système d'équations linéaires de la méthode de Galerkin dans F_h avec la base (5.25) pour le problème

$$\begin{aligned}-\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial y^2}+\frac{\partial u}{\partial x}&=f \text{ dans } D,\\ u&=0 \quad \text{sur} \quad \partial D.\end{aligned} \tag{5.26}$$

Il est connu dès le n° 2.2 que si u^h est cherchée sous la forme

$$u^h=\sum_{(x_i,\ y_j)\in D_h} u_{i,\,j}\omega_{i,\,j}(x,\ y), \tag{5.27}$$

le système s'écrit

$$Au=g, \tag{5.28}$$

avec u et g deux vecteurs de composantes $\{u_{k,\,l}\}$ de (5.27) et $\{g_{k,\,l}\}$ $\left(g_{k,\,l}=\int\limits_D f\omega_{k,\,l}\,dD\right)$ respectivement, les éléments $\{a^{i,\,j}_{k,\,l}\}$ de A étant définis par la formule

$$a^{i,\,j}_{k,\,l}=\int\limits_D\left[\frac{\partial\omega_{i,\,j}}{\partial x}\frac{\partial\omega_{k,\,l}}{\partial x}+\frac{\partial\omega_{i,\,j}}{\partial y}\frac{\partial\omega_{k,\,l}}{\partial y}+\frac{\partial\omega_{i,\,j}}{\partial x}\omega_{k,\,l}\right]dD=$$

$$=\int\limits_D\left[\omega_j\omega_l\frac{\partial\omega_i}{\partial x}\frac{\partial\omega_k}{\partial x}+\omega_i\omega_k\frac{\partial\omega_j}{\partial y}\frac{\partial\omega_l}{\partial y}+\omega_j\frac{\partial\omega_i}{\partial x}\omega_k\omega_l\right]dD, \tag{5.29}$$

où $\omega_i=\omega_i(x)$, $\omega_k=\omega_k(x)$, $\omega_j=\omega_j(y)$, $\omega_l=\omega_l(y)$.

Comme au point précédent, on vérifie sans peine que $a_{k,\,l}^{i,\,j} = 0$ si l'une au moins des inégalités

$$|\,i-k\,|>1, \quad |\,j-l\,|>1$$

a lieu. Il en résulte de suite que A est une matrice tridiagonale par blocs de la forme (5.12). Ci-dessous les valeurs des éléments $\{a_{k,\,l}^{i,\,j}\}$ sous l'hypothèse simplificatrice de réseau régulier de pas h:

$$a_{k,\,l}^{k,\,l} = \frac{1}{h^2}\int\limits_{x_{k-1}}^{x_{k+1}} dx \int\limits_{y_{l-1}}^{y_{l+1}} dy\,[\omega_l^2(y)+\omega_k^2(x)] = \frac{8}{3},$$

$$a_{k,\,l}^{k,\,l-1} = \frac{1}{h^2}\int\limits_{x_{k-1}}^{x_{k+1}} dx \int\limits_{y_{l-1}}^{y_l} dy\,[\omega_{l-1}(y)\,\omega_l(y)-\omega_k^2(x)] = -\frac{1}{3},$$

$$a_{k,\,l}^{k,\,l+1} = -\frac{1}{3},$$

$$a_{k,\,l}^{k-1,\,l} = \int\limits_{x_{k-1}}^{x_k} dx \int\limits_{y_{l-1}}^{y_{l+1}} dy\left[-\frac{\omega_l^2(y)}{h^2}+\frac{\omega_{k-1}(x)\,\omega_k(x)}{h^2}-\right.$$

$$\left.-\frac{1}{h}\,\omega_k(x)\,\omega_l^2(y)\right] = -\frac{1}{3}-\frac{h}{3},$$

$$a_{k,\,l}^{k-1,\,l-1} = \int\limits_{x_{k-1}}^{x_k} dx \int\limits_{y_{l-1}}^{y_l} dy\left[-\frac{1}{h^2}\,\omega_l(y)\,\omega_{l-1}(y)-\right.$$

$$\left.-\frac{1}{h^2}\,\omega_k(x)\,\omega_{k-1}(x)-\omega_k(x)\,\omega_l(y)\,\omega_{l-1}(y)\right] = -\frac{1}{3}-\frac{h}{12}, \qquad (5.30)$$

$$a_{k,\,l}^{k-1,\,l+1} = \int\limits_{x_{k-1}}^{x_k} dx \int\limits_{y_l}^{y_{l+1}} dy\left[-\frac{\omega_{l+1}(y)\,\omega_l(y)}{h^2}-\right.$$

$$\left.-\frac{\omega_{l+1}(x)\,\omega_{k+1}(x)}{h^2}-\frac{\omega_k(x)\,\omega_l(y)\,\omega_{l+1}(y)}{h}\right] = -\frac{1}{3}-\frac{h}{12},$$

$$a_{k,\,l}^{k+1,\,l} = -\frac{1}{3}+\frac{h}{3}, \qquad a_{k,\,l}^{k+1,\,l-1} = -\frac{1}{3}+\frac{h}{12},$$

$$a_{k,\,l}^{k+1,\,l+1} = -\frac{1}{3}+\frac{h}{12}, \qquad (x_k,\ y_l)\in D_h.$$

Il se trouve que le schéma variationnel aux différences obtenu est équivalent pour $f(x, y) \equiv f = \text{const}$ au schéma assez spécial

$$\frac{8}{3}u_{k,l} - \frac{1}{3}u_{k,l-1} - \frac{1}{3}u_{k,l+1} - \left(\frac{1}{3}+\frac{h}{3}\right)u_{k-1,l} -$$
$$-\left(\frac{1}{3}+\frac{h}{12}\right)u_{k-1,l-1} - \left(\frac{1}{3}+\frac{h}{12}\right)u_{k-1,l+1} - \left(\frac{1}{3}-\frac{h}{3}\right)u_{k+1,l} -$$
$$-\left(\frac{1}{3}-\frac{h}{12}\right)u_{k+1,l-1} - \left(\frac{1}{3}-\frac{h}{12}\right)u_{k+1,l+1} = h^2 f. \quad (5.31)$$

On construit facilement (5.31) à l'aide des approximations usuelles à trois points des dérivées secondes à condition d'approcher aux nœuds de D_h la partie différentielle de (5.26):

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \frac{\partial u}{\partial x}\Big|_{\substack{x=x_k \\ y=y_l}} \approx$$
$$\approx \frac{1}{6}\left[\sum_{i=k-1}^{k+1} \beta_{k-i}\left(\frac{\partial^2 u}{\partial y^2}\right)_{\substack{x=x_i \\ y=y_l}} + \sum_{j=l-1}^{l+1} \beta_{l-j}\left(\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x}\right)_{\substack{x=x_k \\ y=y_j}}\right], \quad (5.32)$$

où

$$\beta_{-1} = \beta_1 = 1, \qquad \beta_0 = 4 \quad \text{et} \quad (x_k, y_l) \in D_h.$$

2.5.3. Construction de sous-espaces

La formation de schémas variationnels aux différences pour des équations elliptiques bidimensionnelles se fait moyennant les sous-espaces $F_h \subset \mathring{W}_2^1(D)$ de dimension n. La forme des matrices des équations variationnelles aux différences dépend essentiellement, on l'a vu aux points 2.5.1 et 2.5.2, du choix de F_h et de la triangulation de D.

Nous allons passer en revue des procédés simples de triangulation de D borné de dimension 2 à contour S régulier.

Le domaine discrétisé D_h, réunion d'un nombre fini de triangles disjoints $\Delta_k \subset D$, dont la frontière S_h est linéaire par morceaux, doit respecter en premier lieu les conditions suivantes (voir Oganessian et al. [V; 8]).

1. Les normales à S établissent entre les points de S_h et ceux de S une correspondance biunivoque et la distance entre deux points en correspondance est au plus $\delta_1 h^2$, $\delta_1 > 0$, et ne dépend pas de h.

2. La longueur des côtés et les aires des triangles constitutifs Δ_k sont comprises dans les limites respectives $[l_1 h, l_2 h]$ et $[\gamma_1 h^2, \gamma_2 h^2]$, les constantes positives l_1, l_2, γ_1, γ_2 étant indépendantes de h.

On superpose à D (on le voit sans peine) plusieurs domaines discrets vérifiant ces conditions, si bien qu'on impose à D_h une contrain-

te supplémentaire. Avant de l'énoncer, on rappelle que dans le cas de domaines rectangulaires et de domaines formés de rectangles, les triangulations ont été obtenues de façon naturelle à partir des réseaux rectangulaires (voir points 2.5.1 et 2.5.2). Le problème de construire des domaines discrets de forme déterminée proche de la triangulation des domaines rectangulaires a été discuté par exemple par Godounov, Prokopov [IV; 7]. Quant à la contrainte supplémentaire mentionnée, il est raisonnable de la formuler comme suit.

3. Il existe une bijection continue de D_h sur un rectangle ou un domaine qui ou bien sont limités par des segments parallèles aux axes de coordonnées, ou bien forment avec ces axes un angle de 45°. Cette application est linéaire sur chaque Δ_k et le transforme en un triangle rectangle de côtés h.

Il en résulte par exemple que huit triangles au plus possèdent un sommet en commun.

Décrivons un procédé de triangulation simple. On inclut D dans un rectangle Π et on recouvre Π par un maillage carré de pas h. On déplace les nœuds immédiatement extérieurs à D de façon que la ligne brisée S_h passant par ces points approche bien la frontière S de D (i.e. on est dans la condition 1)). On établit une triangulation des carrés telle qu'on respecte la condition 2). La condition 3) découle de suite du procédé de triangulation (voir fig. 10).

Matsokine [V; 6] a obtenu pour ces triangulations les valeurs des constantes caractéristiques des longueurs et des aires: $l_1 = 0{,}5$; $l_2 = \sqrt{18}/2$; $\gamma_1 = 0{,}125$; $\gamma_2 = 1{,}125$, valeurs indépendantes de la

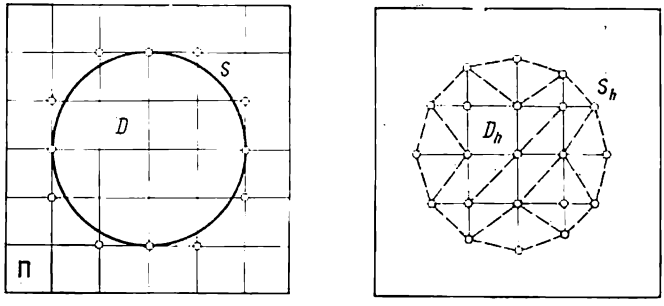


Fig. 10

configuration de D. Quant à la constante δ_1 qui indique le degré d'approximation de la frontière, elle dépend naturellement de la courbure de S.

Il convient de noter que le procédé décrit s'applique aux domaines simplement et multiplement connexes et est facilement généralisable.

S'agissant des domaines simplement connexes, on propose un procédé de formation des domaines discrets qui [illegible] les trois

conditions ci-dessus et qu'une transformation continue linéaire par morceaux change en un rectangle (voir par exemple Diakonov [V ; 2] et Matsokine [V ; 6]).

Pour simplifier, D est supposé convexe. L'idée de la méthode est fort simple. On effectue sur D une transformation continue dont les dérivées présentent des discontinuités à l'intérieur du domaine. On triangule le rectangle obtenu de façon que les côtés des triangles ne coupent pas les lignes de discontinuité des dérivées. La transformation réciproque change ces triangles en des triangles curvilignes qu'on rectifie pour aboutir à D_h cherché.

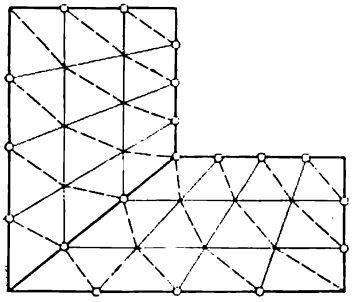


Fig. 11

La première étape s'effectue assez simplement. On entoure par exemple D d'un carré et on l'étire dans le sens des rayons issus du point d'intersection des diagonales du carré. Ces diagonales se confondent avec les lignes de discontinuité mentionnées. Le seul but de cette application est en fait d'établir une loi de génération du domaine discret et on peut parler donc d'une méthode d'application fictive.

La fig. 11 donne cette triangulation pour un domaine en L.

Voyons si ces domaines discrets font l'affaire. On désigne par F_h l'ensemble des fonctions continues linéaires sur chaque triangle élémentaire $\Delta_k \in D_h$ et nulles en dehors de D_h. Il est connu qu'on forme le système d'équations variationnelles aux différences pour le problème aux limites

$$-\Delta u\,(x,\ y) = f\,(x,\ y), \qquad (x,\ y) \in D,$$
$$u\,(x,\ y) = 0, \qquad (x,\ y) \in S$$

moyennant l'équation variationnelle approchée

$$\int\limits_{D_h} \left(\frac{\partial u_h}{\partial x} \frac{\partial v_h}{\partial x} + \frac{\partial u_h}{\partial y} \frac{\partial v_h}{\partial y} \right) dx\,dy = \int\limits_{D_h} f v_h\,dx\,dy$$

pour tous les $v_h \in F_h$,

$u_h \in F_h$ étant la solution approchée cherchée.

La triangulation de D_h est topologiquement équivalente à une triangulation simple d'un domaine formé de rectangles, si bien qu'on numérote en conséquence les sommets des triangles et qu'on définit la base $\{\omega_k\ (x, y)\}$ dans F_h comme au n° 2.5. Dans ce cas, la matrice du système algébrique linéaire obtenu à partir de l'équation varia-

tionnelle approchée possède en principe une structure analogue à celle de la matrice de (5.8).

Les procédés de ce numéro se généralisent au cas de dimension 3.

2.6. Méthodes variationnelles pour des problèmes à plusieurs variables

Nous allons décrire sommairement les procédés de construction de schémas variationnels aux différences pour des problèmes à plusieurs variables.

2.6.1. Construction de sous-espaces

Soit, dans l'espace des variables x, y, z, un domaine borné D de frontière ∂D linéaire par morceaux. La méthode la plus connue pour former $F_h \subset \mathring{W}_2^1(D)$ consiste à trianguler D dans l'espace, i.e. on superpose à D un nombre fini de pyramides triangulaires disjointes Δ_k telles que $D = \bigcup_{k=1}^{N} \Delta_k$. On note h_k la plus grande longueur des arêtes de Δ_k et on forme facilement une suite d'espaces F_h de fonctions morceaux de polynômes définis par le paramètre $h = \max_{1 \leqslant k \leqslant N} h_k$. Ces espaces se définissent comme en uni et bidimensionnel. Illustrons nos paroles sur des approximations linéaires par morceaux et le cube $D = \{x, y, z : 0 < x < 1,\ 0 < y < 1, 0 < z < 1\}$.

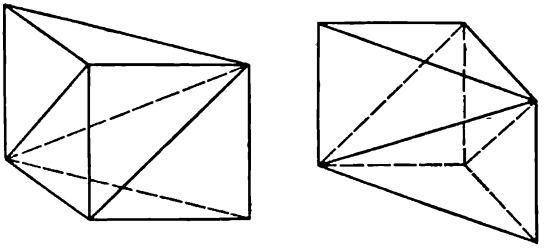


Fig. 12

On divise le segment $[0, 1]$ par les points équidistants $0 = \xi_0 < \xi_1 < \ldots < \xi_{n+1} = 1$, avec $\xi_k = \Delta\xi \cdot k \left(\Delta\xi = \frac{1}{n+1}\right)$, et on couvre D de cubes élémentaires de côté $\Delta\xi$, ce qui exige qu'on trouve les lignes suivant lesquelles D coupe les plans:

$$x = \xi_k, \quad y = \xi_k, \quad z = \xi_k, \quad k = 0, 1, \ldots, n+1.$$

Il suffit maintenant d'effectuer une triangulation d'un cube élémentaire de côté $\Delta\xi$. On procède de différentes manières (on peut agir par exemple comme l'indique la fig. 12: le cube est partagé au préalable en deux prismes).

Ainsi, chaque cube élémentaire est représenté par la réunion de 6 pyramides, la longueur maximale de l'arête étant $h = \sqrt{3}\Delta\xi = \frac{\sqrt{3}}{n+1}$. On traite de même tous les cubes et le domaine D se trouve recouvert complètement par $N = 6\ (n+1)^3$ pyramides. Pour construire F_h, on définit dans chaque Δ_k un polynôme de la forme

$$g_k(x,\ y,\ z) = \sum_{l=0}^{m} \sum_{i_1+i_2+i_3=l} C^k_{i_1,\ i_2,\ i_3} x^{i_1} y^{i_2} z^{i_3} \tag{6.1}$$

tel que la fonction

$$g(x,\ y,\ z) = \begin{cases} g_1(x,\ y,\ z) & \text{si} \quad (x,\ y,\ z) \in \Delta_1, \\ \vdots & \\ g_N(x,\ y,\ z) & \text{si} \quad (x,\ y,\ z) \in \Delta_N \end{cases} \tag{6.2}$$

soit dans $\mathring{W}_2^1(D)$, i.e. elle possède des dérivées premières de carré sommable et devient nulle sur la frontière. La condition est vérifiée dès qu'on a la continuité de g morceaux de polynômes et sa nullité sur ∂D.

S'agissant des fonctions linéaires par morceaux

$$g_k(x,\ y,\ z) = C^k_{0,\ 0,\ 0} + C^k_{1,\ 0,\ 0} x + C^k_{0,\ 1,\ 0} y + C^k_{0,\ 0,\ 1} z, \tag{6.3}$$

$g(x,\ y,\ z)$ est nécessairement continue si les coefficients de g_k se définissent par exemple par les valeurs de g aux sommets de Δ_k correspondant. Autrement dit, il suffit que g_k prenne des valeurs données en ces points.

La méthode est suffisamment universelle et joue pour des domaines limités par des contours de forme plus générale. Ses algorithmes sont par contre très laborieux même pour des domaines simples. Le lecteur trouvera plus loin un autre procédé de formation des F_h qui nous paraît plus commode dans le cas de domaines multidimensionnels de forme spéciale.

Soit, dans l'espace euclidien de dimension p, un domaine D qu'on assimile à la réunion d'un nombre fini de parallélépipèdes rectangles $\{D_\nu\}$ p-dimensionnels et soit $\tilde{D} = \{x_i : a_i \leqslant x_i \leqslant b_i,\ i = 1, \ldots, p\}$ le plus petit parallélépipède qui contient D (sous l'hypothèse de D_ν ayant ses côtés parallèles aux hyperplans de coordonnées). On associe au segment $[a_i,\ b_i]$ de l'axe des x_i pour chaque i, $1 \leqslant i \leqslant p$, le réseau

$$a_i = x_{i,\ 0} < x_{i,\ 1} < \ldots < x_{i,\ N_i} < x_{i,\ N_{i+1}} = b_i, \qquad i = 1, \ldots, p,$$

et on définit le réseau D_h comme étant l'ensemble des points $x_k = (x_{1,\ k_1},\ x_{2,\ k_2},\ \ldots,\ x_{p,\ k_p})$ du domaine D $(1 \leqslant k_i \leqslant N_i,\ i = 1, \ldots, p)$.

Si l'on associe à chaque réseau de dimension 1 le système de fonctions de base à une variable

$$g_{i,k_i}(x_i) = \begin{cases} \dfrac{x_i - x_{i,k_i-1}}{x_{i,k_i} - x_{i,k_i-1}} & \text{si } x_i \in [x_{i,k_i-1},\ x_{i,k_i}], \\ \dfrac{x_i - x_{i,k_i+1}}{x_{i,k_i} - x_{i,k_i+1}} & \text{si } x_i \in [x_{i,k_i},\ x_{i,k_i+1}], \\ 0 \quad \text{dans le cas contraire} \end{cases} \tag{6.4}$$

$$k_i = 1, \ldots, N_i, \quad 1 \leqslant i \leqslant p,$$

on prend le système

$$g_k(x) = \prod_{i=1}^{p} g_{i,k_i}(x_i), \qquad x \in D_h, \quad k = (k_1, k_2, \ldots, k_p), \tag{6.5}$$

pour base de l'espace $F_h \subset \mathring{W}_2^1(D)$ des fonctions linéaires par morceaux. Le cas $p = 2$ nous est suffisamment familier dès le paragraphe précédent.

2.6.2. Formation de schémas variationnels aux différences point par point

Les procédés décrits de génération des F_h et de schémas variationnels aux différences pour les problèmes à plusieurs variables présentent de nombreux points délicats tels que le choix des fonctions de base (premier procédé du point précédent), le calcul des éléments non nuls de la matrice du système, etc. On tourne ces difficultés dans une méthode hybride qui unit l'approche variationnelle à la tactique usuelle par différences finies. Il est vrai que le système d'équations discrétisées n'est plus ni un système aux différences proprement dit, ni un système variationnel aux différences, mais on réussit à affaiblir les impératifs d'une approche par recours à l'autre. Prenons par exemple une équation elliptique de dimension p.

Soit l'équation

$$-\sum_{i=1}^{p} \frac{\partial}{\partial x_i} P_i(x) \frac{\partial u}{\partial x_i} = f \tag{6.6}$$

dans un domaine borné D de frontière ∂D. Considérons pour cette équation le problème de Dirichlet

$$u = 0 \quad \text{sur } \partial D. \tag{6.7}$$

Quand $p \geqslant 2$ et D diffère des domaines étudiés ou d'autres tels qu'on forme sans peine la suite des $F_h \in \mathring{W}_2^1(D)$, on propose le procédé hybride suivant.

Superposons à D un réseau rectangulaire D_h formé par l'intersection de D et de l'ensemble des hyperplans parallèles aux plans de

coordonnées et désignons par $x_k = (x_{1,k_1}, x_{2,k_2}, \ldots, x_{p,k_p})$ les nœuds de D_h. Le système d'équations discrètes sera celui des approximations du problème (6.6), (6.7) associées à tous les nœuds de D_h qui s'obtiendront moyennant des schémas variationnels aux différences de dimension un. On procède comme suit.

Pour approcher le problème proposé au point $x_k \in D_h$, on écrit l'équation sous la forme

$$\sum_{i=1}^{p} \left[-\frac{\partial}{\partial x_i} P_i(x) \frac{\partial u}{\partial x_i} - f_i \right] = 0, \tag{6.8}$$

avec $\sum_{i=1}^{p} f_i = f$. On construit pour chaque terme de la somme (6.8) un schéma variationnel aux différences associé à la droite x_i, $1 \leqslant i \leqslant p$, correspondante et on fait l'addition. Autrement dit, l'équation (6.8) discrétisée au point x_k s'écrit

$$\sum_{i=1}^{p} (L_i^{(h)} u^h - f_i^h) = 0, \tag{6.9}$$

avec $(L_i^{(h)} u^h - f_i^{(h)})$ l'analogue variationnel aux différences du terme

$$-\frac{\partial}{\partial x_i} P_i(x) \frac{\partial u}{\partial x_i} - f_i$$

suivant les segments de l'axe de x_i. Les procédés de construction de schémas variationnels aux différences pour les opérateurs de dimension un ont été examinés en détail au n° 2.3.

Plaçons-nous dans le cas concret du cube unité p-dimensionnel recouvert par un réseau régulier D_h $\left(h = \frac{1}{N+1}\right)$. Selon le numéro 2.4 et (6.9), le système d'équations discrètes pour (6.8) est de la forme

$$\frac{1}{h^2} \sum_{i=1}^{p} [-P_{k_i-1/2} u^h_{k_i-1} + (P_{k_i-1/2} + P_{k_i+1/2}) u^h_{k_i} - \\ - P_{k_i+1/2} u^h_{k_i+1}] - f^h_{k_i} = 0, \qquad x_k \in D_h. \tag{6.10}$$

Nous avons utilisé les notations $(1 \leqslant i \leqslant p)$:

$$u^h_{k_i+\alpha} = u^h(x_{k_1}, \ldots, x_{k_{i-1}}, x_{k_i+\alpha}, x_{k_{i+1}}, \ldots, x_{k_p}),$$
$$\alpha = 0, \pm 1;$$

$$P_{k_i+\alpha} = \frac{1}{h} \int_{x_{k_i+\alpha-1/2}}^{x_{k_i+\alpha+1/2}} P(x_{k_1}, \ldots, x_{k_{i-1}}, x, x_{k_{i+1}}, \ldots, x_{k_p})\, dx, \tag{6.11}$$

$$\alpha = \pm 1/2.$$

2.7. Méthode du domaine auxiliaire

Soit, dans un domaine borné D_1, l'équation différentielle elliptique

$$Lu \equiv -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + c(x)u = f(x), \qquad x = (x_1, \ldots, x_n) \in D_1, \tag{7.1}$$

avec la condition aux limites

$$u(x) = 0, \qquad x \in \partial D_1 \tag{7.2}$$

(problème de Dirichlet).

On suppose que les coefficients et la solution du problème (7.1), (7.2) sont suffisamment réguliers, $a_{ij}(x) = a_{ji}(x)$, $c(x) \geqslant 0$ et on est dans la condition d'ellipticité

$$\inf_{x \in D_1} \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqslant \mu \sum_{i=1}^{n} \xi_i^2, \tag{7.3}$$

où la constante positive μ est indépendante du vecteur arbitraire $\xi = (\xi_1, \ldots, \xi_n)$.

S'il est difficile de programmer la résolution numérique de ce problème par des méthodes des différences ou variationnelles des différences, la cause en est surtout la géométrie de D_1, si bien qu'on ferait mieux de confectionner les programmes pour une classe plus ou moins vaste de domaines. On y arrive par exemple en remplaçant le problème aux limites (7.1), (7.2) par un problème voisin dans un sens, mais donné dans un domaine plus simple, disons dans un parallélépipède. Cette technique proposée par Sauliev [IV; 25] est connue sous le nom de méthode du domaine auxiliaire.

La méthode repose sur ce fait bien connu des physiciens que la densité de substance diffusante varie faiblement dans un milieu à coefficient de diffusion relativement grand. Il est donc naturel qu'on complète le domaine initial D_1 de façon à obtenir un parallélépipède et qu'on prolonge $a_{ij}(x)$ de (7.1) au domaine auxiliaire par des valeurs suffisamment importantes. Si l'on se donne sur le contour du parallélépipède une condition de la forme (7.2), il y a lieu d'espérer

que la solution de ce nouveau problème diffère peut de 0 dans le domaine auxiliaire et coïncide presque partout dans D_1 avec la solution du problème primitif (7.1), (7.2).

Concrétisons ce que nous venons de dire. Soit le problème aux limites de dimension un (voir Sauliev [IV; 25]

$$\frac{d^2u}{dx^2} = -2, \quad 0 < x < 0{,}5, \qquad u(0) = u(0{,}5) = 0, \tag{7.4}$$

dont la solution exacte est $u(x) = x(0{,}5-x)$. Remplaçons (7.4) par

$$\frac{d}{dx} a(x) \frac{dv}{dx} = f(x), \quad 0 < x < 1,$$

$$v(0) = v(1) = 0, \tag{7.5}$$

$$\left(a \frac{dv}{dx}\right)(0{,}5-0) = \left(a \frac{dv}{dx}\right)(0{,}5+0),$$

où

$$a(x) = \begin{cases} 1, & 0 < x \leqslant 0{,}5, \\ \frac{1}{\varepsilon^2}, & 0{,}5 \leqslant x < 1, \end{cases}$$

$$f(x) = \begin{cases} -2, & 0 < x \leqslant 0{,}5, \\ 0, & 0{,}5 \leqslant x < 1. \end{cases}$$

La solution exacte s'écrit dans ce cas

$$v(x) = \begin{cases} x\left(\frac{1+2\varepsilon^2}{1+\varepsilon^2} 0{,}5 - x\right), & 0 \leqslant x \leqslant 0{,}5, \\ \frac{\varepsilon^2}{2(1+\varepsilon^2)}(1-x), & 0{,}5 \leqslant x \leqslant 1. \end{cases}$$

Evidemment $\lim\limits_{\varepsilon \to 0} v(x) = u(x)$ pour $x \in [0;\ 0{,}5]$ et $\lim\limits_{\varepsilon \to 0} v(x) = 0$ pour $x \in (0{,}5;\ 1)$. Le domaine $0{,}5 < x \leqslant 1$ joue donc pour ε suffisamment petits le rôle de région auxiliaire au sens ci-dessus. On obtient des résultats analogues, ou peu s'en faut, dans le cas multidimensionnel.

Décrivons la méthode du domaine auxiliaire pour le problème de Dirichlet (7.1), (7.2). On désigne par D_2 le domaine qu'on ajoute à D_1 pour obtenir le parallélépipède D et par S la frontière commune à D_1 et à D_2.

Soit, dans D, l'équation

$$L_\varepsilon v \equiv -\sum_{i,\,j=1}^{n} \frac{\partial}{\partial x_i} A_{ij}(x) \frac{\partial v}{\partial x_j} + Cv = F(x), \tag{7.6}$$

où

$$A_{ij} = \begin{cases} a_{ij}(x), & x \in D_1, \\ 0, & x \in D_2, \quad i \neq j, \\ \varepsilon^{-2}, & x \in D_2, \quad i = j, \end{cases}$$

$$C(x) = \begin{cases} c(x), & x \in D_1, \\ 0, & x \in D_2, \end{cases}$$

$$F(x) = \begin{cases} f(x), & x \in D_1, \\ 0, & x \in D_2. \end{cases}$$

L'équation (7.6) n'est pas forcément homogène dans D_2. En examinant la torsion d'une barre cylindrique homogène, Konovalov [IV; 9] l'a par exemple choisie non homogène dans D tout entier, ce qui a permis d'interpréter physiquement le problème intervenu.

Posons pour (7.6) le problème aux limites

$$v(x) = 0, \quad x \in \partial D, \tag{7.7}$$

avec les conditions de compatibilité

$$[v(x)]|_S = 0, \quad \left[\sum_{i,\,j=1}^{n} A_{ij} \cos(n, x_i) \frac{\partial v}{\partial x_j}\right]\Big|_S = 0. \tag{7.8}$$

Ici n est la normale à la frontière S; $[\ ]|_S$ désigne le saut de la fonction sur la surface S.

On fait la convention que la solution du problème (7.1), (7.2) est nulle sur D_2 et on évalue la différence $w(x) = u(x) - v(x)$. La fonction $w(x)$ vérifie, on le constate sans peine, le problème aux limites

$$L_\varepsilon w(x) = 0, \quad x \in D, \quad x \notin S, \tag{7.9}$$

$$w(x) = 0, \quad x \in \partial D, \tag{7.10}$$

$$[w(x)]|_S = 0, \quad \left[\sum_{i,\,j=1}^{n} A_{ij}(x) \cos(n, x_i) \frac{\partial w}{\partial x_i}\right]\Big|_S = \varphi(x), \tag{7.11}$$

où $\varphi(x) = \sum_{i,\,j=1}^{n} a_{ij} \cos(n, x_i) \frac{\partial u}{\partial x_i}$, $x \notin S$.

Multiplions (7.9) par $w(x)$ et intégrons sur D compte tenu des conditions (7.10), (7.11), il vient l'identité

$$\int\limits_{D_1} \Big(\sum_{i,j=1}^{n} a_{ij} \frac{\partial w}{\partial x_i} \frac{\partial w}{\partial x_j} + Cw^2 \Big) dx + \frac{1}{\varepsilon^2} \int\limits_{D_2} \sum_{i=1}^{n} \Big(\frac{\partial w}{\partial x_i} \Big)^2 dx = \\ = - \int\limits_{S} \varphi(x)\, w(x)\, ds. \qquad (7.12)$$

Rejetons le premier terme de la somme à gauche et appliquons au second membre l'inégalité de Cauchy-Schwarz :

$$\frac{1}{\varepsilon^2} \int\limits_{D_2} \sum_{i=1}^{n} \Big(\frac{\partial w}{\partial x_i} \Big)^2 dx \leqslant \sqrt{\int\limits_{S} \varphi^2(x)\, ds} \sqrt{\int\limits_{S} w^2(x)\, ds}. \qquad (7.13)$$

Le second membre s'évalue moyennant l'inégalité

$$\int\limits_{S} w^2(x)\, ds \leqslant C_1 \Big(\delta \int\limits_{\omega_\delta} w^2(x)\, dx + \frac{1}{\delta} \int\limits_{\omega_\delta} \sum_{i=1}^{n} \Big(\frac{\partial w}{\partial x_i} \Big)^2 dx \Big), \qquad (7.14)$$

juste pour la bande limitrophe ω_δ de largeur $0 < \delta < \delta_0$, δ_0 étant indépendant de $w(x)$ (voir Oganessian, Rivkind, Roukhovetz [V ; 8]), et l'inégalité de Friedrichs

$$\int\limits_{D} w^2(x)\, dx \leqslant C_2 \int\limits_{D} \sum_{i=1}^{n} \Big(\frac{\partial w}{\partial x_i} \Big)^2 dx, \qquad (7.15)$$

vérifiée par $w(x)$ nulles sur une portion de frontière de D.

Faisons $\delta = \delta_0$ dans (7.14) et étendons le domaine d'intégration au second membre, on a

$$\int\limits_{S} |w(x)|^2\, ds \leqslant C_3 \Big(\int\limits_{D_2} |w(x)|^2\, dx + \int\limits_{D_2} \sum_{i=1}^{n} \Big| \frac{\partial w}{\partial x_i} \Big|^2 dx \Big). \qquad (7.16)$$

Comme $w(x) = 0$, $x \in \partial D$, on applique (7.15) au premier terme de la somme :

$$\int\limits_{S} |w(x)|^2\, ds \leqslant C_4 \int\limits_{D_2} \sum_{i=1}^{n} \Big| \frac{\partial w}{\partial x_i} \Big|^2 dx. \qquad (7.17)$$

On prend en considération la dernière inégalité et on trouve à partir de (7.13)

$$\Big(\int\limits_{D_2} \sum_{i=1}^{n} \Big(\frac{\partial w}{\partial x_i} \Big)^2 dx \Big)^{1/2} \leqslant C_4 \varepsilon^2, \qquad (7.18)$$

et à partir de (7.15)

$$\int_{D_2} |w(x)|^2 dx \leqslant C_5 \varepsilon^2. \tag{7.19}$$

L'identité (7.12) entraîne de même à l'aide de la condition d'ellipticité (7.3) et de la non-négativité de $f(x)$:

$$\mu \int_{D_1} \sum_{i=1}^{n} \left(\frac{\partial w}{\partial x_i}\right)^2 dx \leqslant \sqrt{\int_S |\varphi|^2 ds} \sqrt{\int_S |w(x)|^2 ds} \leqslant C_4 \varepsilon^2. \tag{7.20}$$

L'inégalité de Friedrichs généralisée

$$\int_{D_1} |w(x)|^2 dx \leqslant C_6 \left(\int_S |w(x)|^2 ds + \int_{D_1} \sum_{i=1}^{n} \left(\frac{\partial w}{\partial x_i}\right)^2 dx \right) \tag{7.21}$$

jointe aux estimations (7.17), (7.18) et (7.20) implique de plus

$$\int_{D_1} |w(x)|^2 dx \leqslant C_7 \varepsilon^2. \tag{7.22}$$

Ainsi, la solution $v(x)$ du problème (7.6), (7.7), (7.8) approche la solution $u(x)$ de (7.1), (7.2) dans $W_2^1(D_1)$ à ε près, i.e. (voir (7.20) et (7.22))

$$\|u - v\|_{W_2^1(D_1)} \leqslant C_8 \varepsilon, \tag{7.23}$$

la constante C_8 étant indépendante de ε.

Koptchénov [IV; 10] a amélioré au mieux (7.23):

$$\| u - v \|_{C(D_1)} \leqslant C_9 \varepsilon^2. \tag{7.24}$$

La résolution approchée du problème de Dirichlet (7.1), (7.2) dans un domaine borné D_1 quelconque s'effectue donc selon le schéma suivant. Construire le plus petit parallélépipède D renfermant D_1. Choisir ε tel que la solution du problème (7.6), (7.7), (7.8) approche celle du problème proposé avec une précision désirée. Résoudre le premier problème par une méthode des différences avec le degré d'exactitude voulu. Pour la convergence des solutions aux différences vers la solution exacte, consulter par exemple Rivkind [IV; 22].

Utilisons enfin l'idée de la méthode du domaine auxiliaire pour le troisième problème aux limites

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial u}{\partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x) u = f(x), \tag{7.25}$$

$$x = (x_1, \ldots, x_n) \in D_1,$$

$$\sum_{i,j=1}^{n} a_{ij}(x) \cos(n, x_i) \frac{\partial u}{\partial x_j} + \sigma(x) u = 0, \qquad x \in \partial D_1, \tag{7.26}$$

la fonction $\sigma(x)$ étant suffisamment régulière et non négative sur ∂D_1.

Complétons le domaine D_1 de façon à obtenir un parallélépipède $D \supset D_1$. On suppose l'existence d'une sphère de rayon $\rho > 0$ intérieure à D_1 telle qu'elle touche n'importe quel point de la surface ∂D_1. On définit dans une bande limitrophe de largeur $\varepsilon \leqslant \rho$ de D_1 une fonction $\sigma_\varepsilon(x)$ par la formule

$$\sigma_\varepsilon(x) = \frac{2}{\varepsilon}\sigma(x_\tau)\left(1+\frac{\tau}{\varepsilon}\right),$$
$$-\varepsilon \leqslant \tau \leqslant 0, \quad x_{\tau i} = x_i - \tau\cos(n, x_i),$$

où $|\tau|$ est la distance comptée suivant la normale entre x et ∂D_1. On pose $\tau_\varepsilon(x) = 0$ partout ailleurs dans D. Soit, dans D, l'équation

$$-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}A_{ij}(x)\frac{\partial v}{\partial x_j}+\sum_{i=1}^{n}B_i(x)\frac{\partial v}{\partial x_i}+C(x)v+\sigma_\varepsilon(x)v=F(x), \tag{7.27}$$

où

$$A_{ij}(x)=\begin{cases} a_{ij}(x), & x\in D_1,\\ 0, & x\in D_2, \quad i\neq j,\\ \varepsilon, & x\in D_2, \quad i=j;\end{cases}$$

$$B_i(x)=\begin{cases} b_i(x), & x\in D_1,\\ 0, & x\in D_2;\end{cases}$$

$$C(x)=\begin{cases} c(x), & x\in D_1,\\ 0, & x\in D_2;\end{cases}$$

$$F(x)=\begin{cases} f(x), & x\in D_1,\\ 0, & x\in D_2.\end{cases}$$

On pose pour (7.27) le problème aux limites

$$v(x) = 0, \quad x \in \partial D, \tag{7.28}$$

avec les conditions de compatibilité

$$[v(x)]\,|_{\partial D_1} = 0, \quad \left[\sum_{i,j=1}^{n}A_{ij}(x)\cos(n, x_i)\frac{\partial v}{\partial x_j}\right]\Big|_{\partial D_1} = 0. \tag{7.29}$$

Sous certaines hypothèses sur la régularité des coefficients du problème (7.25), (7.26), Roukhovetz ([IV; 23] démontre la majoration

$$\|u-v\|^2_{W_2^1(D_1)} \leqslant \varepsilon C_{10}\|f\|^2_{L_2(D_1)}, \tag{7.30}$$

avec une constante C_{10} indépendante de $\varepsilon > 0$.

Le passage du troisième problème aux limites (7.25), (7.26) au problème de Dirichlet (7.27), (7.28), (7.29) peut être expliqué comme suit. Supposons que la solution de celui-ci converge, lorsque $\varepsilon \to 0$, vers une fonction $\tilde{u}(x)$ en norme de $W_2^1(D)$. La solution $v(x)$ de ce problème vérifie l'identité intégrale

$$\int_{D_1} \sum a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial \Phi}{\partial x_j} dx + \varepsilon \int_{D_2} \sum \frac{\partial v}{\partial x_i} \frac{\partial \Phi}{\partial x_i} dx + \\ + \int_{D_1} C(x) v \Phi \, dx + \int_{D_1} \sigma_\varepsilon v \Phi \, dx = \int_{D_1} f \Phi \, dx \quad (7.31)$$

quelle que soit $\Phi \in \mathring{W}_2^1(D_1)$ (pour plus de simplicité on pose $b_i(x) = 0$). Faisons tendre ε de (7.31) vers 0, il vient l'identité

$$\int_{D_1} \sum a_{ij} \frac{\partial \tilde{u}}{\partial x_i} \frac{\partial \Phi}{\partial x_j} dx + \int_{D_1} C \tilde{u} \Phi \, dx + \int \sigma \tilde{u} \Phi \, ds = \int_{D_1} f \Phi \, dx \quad (7.32)$$

juste pour toutes les $\Phi \in W_2^1(D_1)$. Du moment que la solution du problème (7.25), (7.26) satisfait à (7.32), il y a coïncidence de $u(x)$ et $\tilde{u}(x)$ dans D_1.

2.8. Problèmes d'extrémum avec contraintes et inéquations variationnelles

Nous avons étudié dans ce chapitre diverses positions variationnelles de certains problèmes de la physique mathématique et des méthodes d'approximation pour les problèmes d'extrémum associés. Cette analyse avait ses particularités : on se plaçait dans l'espace hilbertien tout entier sans aucune contrainte supplémentaire sur la solution cherchée. Mais de nombreux problèmes intervenant en physique, mécanique, géophysique, commande optimale et économie conduisent à des problèmes d'extrémum dont on cherche la solution dans une classe fonctionnelle plus restreinte que le domaine de définition des fonctionnelles correspondantes ne le permet. On aborde de tels problèmes avec différentes optiques et on en décrit la résolution dans Lions [V], Duvaut et Lions [V], Stampacchia [V], Brézis et Stampacchia [V], Tchernoousko, Banitchouk [III; 16], Bensoussan, Lions et Temam [III], Glowinski [V], Glowinski, Lions et Trémolières [V].

On donnera plus loin des renseignements sur l'existence et l'unicité pour des problèmes d'extrémum avec contraintes et la caractérisation des solutions par des inéquations variationnelles. On utilisera des inéquations variationnelles pour analyser les propriétés des solutions et on discutera divers moyens d'approximation.

2.8.1. Eléments de la théorie générale

Soit U un espace de Hilbert sur le corps des nombres réels muni du produit scalaire $(,)_U$ et de la norme associée $\| \cdot \|_U$. Supposons qu'on a défini sur U une forme bilinéaire symétrique continue $\pi(u, v)$, une forme linéaire continue $L(v)$ et un ensemble $U_d \subseteq U$ fermé convexe. On introduit la fonctionnelle quadratique

$$J(v) = \pi(v, v) - 2L(v), \tag{8.1}$$

la fonctionnelle $\pi(v, v)$ étant supposée définie positive dans U, i e. on trouve une constante $\alpha > 0$ telle que

$$\pi(v, v) \geqslant \alpha \| v \|_U^2 \tag{8.2}$$

quel que soit $v \in U$. On a l'affirmation suivante démontrée dans la monographie de Lions [II].

Sous les hypothèses faites, il existe un élément u et un seul dans U_d qui est solution du problème

$$J(u) = \inf_{v \in U_d} J(v). \tag{8.3}$$

Un moyen universel et bien élaboré de donner un sens intuitif au problème d'extrémum (8.3) afin d'étudier les propriétés de sa solution, consiste à utiliser des inéquations variationnelles. Il s'appuie sur le résultat suivant.

Avec les hypothèses faites, l'élément $u \in U_d$ est solution de (8.3) si et seulement si tout $v \in U_d$ vérifie

$$\pi(u, v - u) \geqslant L(v - u). \tag{8.4}$$

Démontrons (8.4) dans un sens. On suppose que u est solution de (8.3), auquel cas tout $v \in U_d$ et tout $\Theta \in (0, 1)$ satisfont à

$$J(u) \leqslant J((1 - \Theta)u + \Theta v) \tag{8.5}$$

(nous utilisons la convexité de U_d, i.e. le fait que $\varphi = (1 - \Theta)w + \Theta v$ appartient à U_d quels que soient $w, v \in U_d$ et $\Theta \in (0, 1)$). D'où

$$\frac{J(u + \Theta(v - u)) - J(u)}{\Theta} \geqslant 0 \tag{8.6}$$

ou, par passage à la limite pour $\Theta \to 0$ et compte tenu de la forme de J:

$$\lim_{\Theta \to 0} \frac{J(u + \Theta(v - u)) - J(u)}{\Theta} =$$

$$= \lim_{\Theta \to 0} \{2[\pi(u, v - u) - L(v - u)] + \Theta \pi(v - u, v - u)\} =$$

$$= 2[\pi(u, v - u) - L(v - u)] \geqslant 0 \tag{8.7}$$

quel que soit $v \in U_d$, ce qui prouve (8.4).

Admettons que l'élément u vérifie (8.4). La fonctionnelle quadratique $J(v)$ est convexe, i.e. n'importe quels $v, w \in U$ et $\Theta \in (0, 1)$ respectent l'inégalité

$$J((1-\Theta)w + \Theta v) \leqslant (1-\Theta) J(w) + \Theta J(v)$$

ou, ce qui revient au même,

$$J(v) - J(w) \geqslant \frac{J((1-\Theta)w + \Theta v) - J(w)}{\Theta}. \tag{8.8}$$

Substituons u à w, il vient

$$J(v) - J(u) \geqslant \frac{J((1-\Theta)u + \Theta v) - J(u)}{\Theta}.$$

Faisons tendre Θ vers 0. On a par suite de (8.7) et (8.4)

$$J(v) - J(u) \geqslant 2\,[\pi(u, v-u) - L(v-u)] \geqslant 0,$$

donc

$$J(u) \leqslant J(v),$$

quel que soit $v \in U_d$, i.e. u est solution du problème (8.3).

Appliquons ce résultat à deux cas concrets. Soit U_d un sous-espace de U qui est fermé pour la norme de U. Posons dans (8.4) $v = u \pm \varphi$, φ étant un élément quelconque de U_d. On aboutit à deux inégalités :

$$\pi(u, \varphi) \geqslant L(\varphi),$$
$$-\pi(u, \varphi) \geqslant -L(\varphi).$$

Il en résulte qu'avec l'hypothèse faite, l'élément u est solution du problème (8.3) si et seulement si tout $\varphi \in U$ vérifie l'égalité

$$\pi(u, \varphi) = L(\varphi). \tag{8.9}$$

Cette relation n'est rien d'autre que l'équation d'Euler pour le problème variationnel

$$J(u) = \inf_{v \in U_d} J(v). \tag{8.10}$$

Plaçons-nous dans le second cas et supposons U_d être un cône. (On appelle cône un ensemble fermé convexe $K \subset U$ qui, s'il contient un φ, contient aussi $\gamma\varphi$ quel que soit $\gamma \geqslant 0$, et tel que si φ, $-\varphi \in K$, alors φ est élément nul de U.) On porte dans (8.4) $v+u$ à la place de v et on obtient

$$\pi(u, v) \geqslant L(v), \tag{8.11}$$

inégalité vérifiée par tout $v \in U_d$. Nous avons utilisé la propriété suivante de U_d : si U_d contient deux vecteurs v et w quelconques,

alors $v+w \in U_d$. En posant $v = 0$ dans (8.4) et $v = u$ dans (8.11), on trouve respectivement

$$-\pi(u, u) \geqslant -L(u),$$
$$\pi(u, u) \geqslant L(u),$$

qui entraînent de suite

$$\pi(u, u) = L(u). \tag{8.12}$$

On montre l'équivalence entre (8.11)-(8.12) et (8.4). Il suffit d'établir que (8.4) découle de (8.11)-(8.12). En effet, on retranche (8.12) de (8.11):

$$\pi(u, v) - \pi(u, u) \geqslant L(v - L(u),$$

et cette différence conduit, en vertu de la bilinéarité de π et de la linéarité de L, à

$$\pi(u, v-u) \geqslant L(v-u),$$

juste pour tout $v \in U_d$. Ainsi, (8.4) et (8.11)-(8.12) sont équivalents si U_d est un cône.

2.8.2. Quelques problèmes d'extrémum

Nous nous bornerons au cas unidimensionnel où les éléments $u \in U$ sont des fonctions de la variable x et U constitue l'espace de Sobolev $W_2^1(0, 1)$, car ce cas nous paraît le plus apte à illustrer les résultats précédents. L'espace U est muni d'un produit scalaire et d'une norme définis par

$$(u, v)_U = \int_0^1 \left[\frac{du}{dx}\frac{dv}{dx} + uv\right] dx,$$
$$\|u\|_U = (u, u)_U^{1/2}. \tag{8.13}$$

Introduisons la fonctionnelle bilinéaire

$$\pi(u, v) = (u, v)_U \tag{8.14}$$

et la fonctionnelle linéaire

$$L(v) = \int_0^1 fv\,dx, \tag{8.15}$$

avec $f = f(x)$ un élément fixe de $L_2(0, 1)$. Evidemment $\pi(u, v)$ est définie positive symétrique continue et $L(v)$ continue. Ces propriétés découlent de

$$\pi(u, v) \leqslant \|u\|_U \cdot \|v\|_U,$$
$$\pi(v, v) = \|v\|_U^2, \tag{8.16}$$

$$L(v) \leqslant \|f\|_{L_2(0, 1)} \cdot \|v\|_{L_2(0, 1)} \leqslant \|f\|_{L_2(0, 1)} \cdot \|v\|_U. \tag{8.17}$$

Ainsi, la fonctionnelle quadratique

$$J(v) = (v, v)_U - 2\int_0^1 fv\,dx \tag{8.18}$$

respecte toutes les conditions du point précédent, si bien que l'élément $u \in U_d$ solution du problème d'extrémum

$$J(u) = \inf_{v \in U_d} J(v), \tag{8.19}$$

existe et est unique si $U_d \subseteq U$ est un ensemble convexe fermé. Cet élément est défini de façon unique par l'inégalité

$$\pi(u, v-u) \geqslant L(v-u), \tag{8.20}$$

quel que soit $v \in U_d$. Il y a lieu de noter que la solution $u(x)$ du problème (8.19) est nécessairement une fonction continue parce que toute $v(x) \in W_2^1(0, 1)$ se trouve dans l'espace de Banach $C(0, 1)$ des fonctions continues (voir Sobolev [I ; 6]). Il existe de plus une constante α telle qu'on ait pour toute fonction $v(x) \in U = W_2^1(0, 1)$

$$\|v\|_{C(0,1)} \leqslant \alpha \|v\|_U, \tag{8.21}$$

i.e. la convergence de la suite $\{v_k(x)\}$ vers $v^\infty(x)$ en norme de $W_2^1(0, 2)$ entraîne la même propriété pour la norme de $C(0, 1)$. Concrétisons nos résultats.

Soit $U_d = U$, auquel cas U_d est un sous-espace propre de U (qui coïncide avec U) et la solution $u(x)$ du problème (8.19) vérifie pour toute $v(x) \in U$ l'égalité

$$(u, v)_U = L(v)$$

(premier exemple du point 2.8.1). On en déduit par des raisonnements standard (voir Mikhlin [I]) que $u(x)$ est solution généralisée du problème de Neumann

$$\begin{aligned} -\frac{d^2u}{dx^2} + u &= f(x), \quad 0 < x < 1, \\ \frac{du}{dx} &= 0 \quad \text{pour} \quad x = 0, \\ \frac{du}{dx} &= 0 \quad \text{pour} \quad x = 1. \end{aligned} \tag{8.22}$$

Si $f(x)$ est par exemple continue, alors $u(x)$ constitue solution classique de ce problème.

Admettons que $U_d = \mathring{W}_2^1(0, 1)$, i.e. il est l'ensemble des fonctions de $W_2^1(0, 1)$ qui s'annulent aux extrémités de l'intervalle $(0, 1)$. Il est connu que $\mathring{W}_2^1(0, 1)$ est un sous-espace fermé de $W_2^1(0, 1)$ (car c'est un espace complet pour la métrique $W_2^1(0, 1)$). Aussi la solu-

tion $u(x) \in \mathring{W}_2^1(0, 1)$ du problème (8.19) satisfait pour toute $v(x) \in \in \mathring{W}_2^1(0, 1)$ à l'égalité

$$(u, v)_U = L(v)$$

(premier exemple du point 2.8.1). Comme dans l'exemple précédent, on en déduit que $u(x)$ est solution généralisée du problème de Dirichlet

$$-\frac{d^2u}{dx^2} + u = f(x), \quad 0 < x < 1, \qquad u(0) = u(1) = 0. \tag{8.23}$$

Considérons deux cas où U_d n'est pas une partie de U, i.e. sa structure est plus compliquée. On définit U_d par la relation

$$U_d = \{u(x) \mid u(x) \in \mathring{W}_2^1(0, 1), \quad u(x) \geqslant 0 \quad \text{pour} \quad x \in (0, 1)\}, \tag{8.24}$$

i.e. on est en présence d'un ensemble de fonctions $\in W_2^1(0, 1)$ non négatives sur $(0, 1)$ et nulles aux extrémités de cet intervalle. On montre que U_d est un cône de $W_2^1(0, 1)$. Cela est vrai si les conditions suivantes se trouvent satisfaites : U_d est convexe ; U_d est fermé ; si $v(x) \in U_d$, alors $\gamma v(x) \in U_d$ quel que soit $\gamma \geqslant 0$; si $v(x)$, $-v(x) \in U_d$, alors $v(x) \equiv 0$.

La condition de convexité signifie que si U_d contient $u(x)$ et $v(x)$, il contient également $w(x) = (1 - \Theta) u(x) + \Theta v(x)$ quel que soit $\Theta \in (0, 1)$. Elle s'avère valable en raison de la non-négativité de $w(x)$ qui est en effet une combinaison linéaire à coefficients positifs de fonctions non négatives.

Soit $v^k(x)$, $k = 0, 1, \ldots$, une suite de fonctions de U_d qui tend pour la norme de $\mathring{W}_2^1(0, 1)$ vers une fonction $w^\infty(x) \in \mathring{W}_2^1(0, 1)$. Comme $v^\infty(x) \in C(0, 1)$, la suite $\{v^k(x)\}$ converge, selon (2.21), vers $v^\infty(0, 1)$ pour la norme de $C(0, 1)$. Ainsi, $v^\infty(x)$ est non négative, nulle aux extrémités de l'intervalle $(0, 1)$ et elle appartient donc à U_d (on est dans la deuxième condition).

Les deux dernières propriétés sont établies de façon triviale.

Etant donné U_d défini par (8.24), le problème (8.19) admet donc, d'après le point 2.8.1, une solution $u(x)$ et une seule si et seulement si toute $v(x) \in U_d$ vérifie l'inégalité

$$\pi(u, v) \geqslant L(v) \tag{8.25}$$

et

$$\pi(u, u) = L(u). \tag{8.26}$$

Etablissons les propriétés de $u(x)$.

Supposons que $u(x)$ est positive en un point $\xi \in (0, 1)$, auquel cas sa continuité entraîne l'existence d'un intervalle $(\xi^-, \xi^+) \subseteq (0, 1)$ tel que $\xi \in (\xi^-, \xi^+)$, $u(x)$ soit positive sur cet intervalle et $u(\xi^-) = u(\xi^+) = 0$ (il se peut en particulier que $\xi^- = 0$ et $\xi^+ = 1$). Désignons par $D(\xi^-, \xi^+)$ l'ensemble des fonctions indéfiniment dérivables à support borné dans (ξ^-, ξ^+) et identiquement nulles partout ailleurs. On sait (voir Sobolev [I ; 6]) que $D(\xi^-, \xi^+)$ est dense dans $\mathring{W}_2^1(\xi^-, \xi^+)$ et que toute fonction $\varphi(x)$ de cet ensemble jouit de la propriété : avec tout $\varepsilon > 0$ suffisamment petit, les fonctions

$$v_\varepsilon(x) = v(x) \pm \varepsilon\varphi(x) \tag{8.27}$$

sont dans U_d. Portons $v(x) = v_\varepsilon(x)$ dans (8.25), il vient

$$\pi(u, u) \pm \varepsilon\pi(u, \varphi) \geqslant L(u) \pm \varepsilon L(\varphi). \tag{8.28}$$

D'où, compte tenu de (8.26),

$$\pi(u, \varphi) \geqslant L(\varphi), \tag{8.29}$$

$$-\pi(u, \varphi) \geqslant -L(\varphi), \tag{8.30}$$

donc

$$\pi(u, \varphi) = L(\varphi), \tag{8.31}$$

quel que soit $\varphi \in D(\xi^-, \xi^+)$.

Etant donné $\varphi(x) \equiv 0$ pour $x \notin (\xi^-, \xi^+)$, on met (8.31) sous la forme

$$\int_{\xi^-}^{\xi^+} \left[\frac{du}{dx}\frac{d\varphi}{dx} + u\varphi\right] dx = \int_{\xi^-}^{\xi^+} f\varphi\, dx, \tag{8.32}$$

avec $\varphi(x)$ une fonction quelconque de $D(\xi^-, \xi^+)$ et $u(x) \in \mathring{W}_2^1(\xi^-, \xi^+)$. L'ensemble $D(\xi^-, \xi^+)$ est dense dans $\mathring{W}_2^1(\xi^-, \xi^+)$, si bien que (8.32) a également lieu pour tout $\varphi \in \mathring{W}_2^1(\xi^-, \xi^+)$. D'où, selon le deuxième exemple ci-dessus, la propriété, pour $u(x)$, d'être sur le segment $[\xi^-, \xi^+]$ une solution généralisée du problème de Dirichlet

$$\begin{gathered} -\frac{d^2u}{dx^2} + u = f(x), \qquad \xi^- < x < \xi^+, \\ u(\xi^-) = u(\xi^+) = 0. \end{gathered} \tag{8.33}$$

Ainsi, on associe à tout point $\xi \in (0, 1)$ en lequel la solution $u(x)$ de (8.19) est positive, un intervalle (ξ^-, ξ^+) où $u(x)$ est solution généralisée du problème (8.33).

Faisons le bilan : l'intervalle $\Omega = (0, 1)$ est représenté par

$$\Omega = \Omega^0 \cup \Omega^+, \tag{8.34}$$

où Ω^+ est un ensemble ouvert, réunion d'un certain nombre (peut-être dénombrable) d'intervalles sur lesquels $u(x)$ est positive et constitue une solution généralisée d'un problème du type (8.33), et Ω^0 le complémentaire (peut-être vide) de Ω^+ par rapport à Ω sur lequel $u(x) \equiv 0$.

Le problème considéré est de classe de problèmes à frontière libre (voir Lions [V]) qui interviennent par exemple en mécanique des milieux continus. Les inconnues sont aussi bien les bornes des intervalles constitutifs de Ω^+ que la fonction $u(x)$ sur ces intervalles.

Plaçons-nous enfin dans un cas intéressant où U_d n'est ni une partie ni un cône de U. On suppose données sur $(0, 1)$ deux fonctions $F_1(x)$ et $F_2(x)$ de $W_2^1(0, 1)$ qui vérifient pour tout $x \in (0, 1)$ l'inégalité

$$F_1(x) < F_2(x) \tag{8.35}$$

et

$$\begin{aligned} &F_1(0) \leqslant 0 \leqslant F_2(0), \\ &F_1(1) \leqslant 0 \leqslant F_2(1). \end{aligned} \tag{8.36}$$

On définit alors U_d par la relation

$$U_d = \{u(x) \mid u(x) \in \overset{\circ}{W}{}_2^1(0, 1),\ F_1(x) \leqslant u(x) \leqslant F_2(x) \text{ quel que soit } x \in (0, 1)\}, \tag{8.37}$$

i.e. il s'agit d'un ensemble des fonctions de $W_2^1(0, 1)$ qui s'annulent aux extrémités de $(0, 1)$ et vérifient sur cet intervalle l'inégalité

$$F_1(x) \leqslant u(x) \leqslant F_2(x). \tag{8.38}$$

On procède comme plus haut et on montre que U_d est un ensemble convexe fermé de $W_2^1(0, 1)$. On en tire, compte tenu du point 2.8.1, l'unicité pour le problème (8.19) avec U_d ainsi choisi.

Admettons qu'en un point $\xi \in (0, 1)$

$$F_1(\xi) < u(\xi) < F_2(\xi). \tag{8.39}$$

La fonction $u(x)$ étant continue, on trouve des points $\xi^-, \xi^+ \in [0, 1]$ tels que $\xi \in (\xi^-, \xi^+)$ et la fonction $u(x)$ satisfasse sur (ξ^-, ξ^+) aux inégalités

$$F_1(x) < u(x) < F_2(x), \tag{8.40}$$

et aux extrémités de cet intervalle aux conditions aux limites

$$\begin{aligned} u(\xi^-) &= \begin{cases} \text{ou bien } F_1(\xi^-), \\ \text{ou bien } F_2(\xi^-), \end{cases} \\ u(\xi^+) &= \begin{cases} \text{ou bien } F_1(\xi^+), \\ \text{ou bien } F_2(\xi^+). \end{cases} \end{aligned} \tag{8.41}$$

Ainsi, on représente l'intervalle $\Omega = (0, 1)$ par

$$\Omega = \Omega^0 \cup \Omega^+, \tag{8.42}$$

où Ω^+ est la réunion des intervalles sur lesquels $u(x)$ vérifie (8.40) et Ω^0 le complémentaire de Ω^+ par rapport à Ω sur lequel $u(x)$ est égale ou bien à $F_1(x)$, ou bien à $F_2(x)$. Si $F_1(x)$ et $F_2(x)$ ne s'annulent pas aux extrémités de $(0, 1)$, alors Ω^0 constitue un ensemble fermé, i.e. il est la réunion de segments et de points isolés de $(0, 1)$. Ceci étant, la fonction $u(x)$ est égale ou bien à $F_1(x)$, ou bien à $F_2(x)$ sur chaque segment de la dernière réunion.

Voyons le comportement de $u(x)$ sur un intervalle $(\xi^-, \xi^+) \subset \Omega^+$. On sait qu'avec tout $\varepsilon > 0$ suffisamment petit et toute $\varphi(x) \in D(\xi^-, \xi^+)$, chaque fois que $u(x)$ appartient à U_d, les fonctions $v_\varepsilon(x) = u(x) \pm \varepsilon\varphi(x)$ lui appartiennent aussi, si bien que la substitution $v_\varepsilon(x) = v(x)$ dans (8.20) donne

$$\pi(u, \varphi) \geqslant L(\varphi),$$
$$-\pi(u, \varphi) \geqslant -L(\varphi).$$

On en déduit pour toute $\varphi(x) \in D(\xi^-, \xi^+)$ l'égalité

$$\pi(u, \varphi) = L(\varphi) \tag{8.43}$$

ou, ce qui revient au même,

$$\int_{\xi^-}^{\xi^+} \left[\frac{du}{dx}\frac{d\varphi}{dx} + u\varphi\right] dx = \int_{\xi^-}^{\xi^+} f\varphi\, dx, \tag{8.44}$$

avec $u(x)$ respectant les conditions aux limites

$$\begin{aligned} u(\xi^-) &= g^-, \\ u(\xi^+) &= g^+, \end{aligned} \tag{8.45}$$

et g^-, g^+ définis par les relations

$$\begin{aligned} g^- &= \begin{cases} \text{ou bien } F_1(\xi^-), \\ \text{ou bien } F_2(\xi^-), \end{cases} \\ g^+ &= \begin{cases} \text{ou bien } F_1(\xi^+), \\ \text{ou bien } F_2(\xi^+). \end{cases} \end{aligned} \tag{8.46}$$

Les exemples antérieurs autorisent à dire que sur chaque (ξ^-, ξ^+) la fonction $u(x)$ est solution généralisée du problème de Dirichlet

$$\begin{aligned} -\frac{d^2u}{dx^2} + u &= f(x), \quad \xi^- < x < \xi^+, \\ u(\xi^-) &= g^-, \\ u(\xi^+) &= g^+. \end{aligned} \tag{8.47}$$

Ainsi, le problème énoncé consiste non seulement à chercher les bornes des intervalles de Ω^+ et à calculer $u(x)$ sur ces intervalles (comme dans le problème précédent), mais aussi à définir les valeurs de $u(x)$ aux extrémités de ces intervalles. Nous sommes en présence d'un autre problème à frontière libre.

2.8.3. Méthodes numériques pour les problèmes d'extrémum

Selon les propriétés de la fonctionnelle quadratique $J(v)$ et de l'ensemble $U_d \subseteq U$ réalisant l'extrémum cherché, on emploie telle ou telle méthode de résolution. S'agissant des deux premiers exemples ci-dessus où la résolution approchée du problème d'extrémum se ramène à la recherche de la solution classique ou de la solution généralisée du problème aux limites linéaire pour l'équation différentielle, il y a intérêt à aborder le problème aux limites directement obtenu. Dans la pratique, cette tactique s'avère très laborieuse dans les cas 3 et 4 où l'on ignore aussi bien les frontières des domaines qui renferment la solution que les conditions aux limites. Ces problèmes sont tout de même justiciables de nombreux procédés qui utilisent par exemple la régularisation, la pénalisation, les idées de la programmation non linéaire (voir Lions [V]). Nous allons, pour notre part, réduire le problème (8.3) à un problème en dimension finie et chercher l'extrémum d'une fonction quadratique de n variables sur un ensemble convexe borné d'un espace de dimension n.

Soit U^h une suite de sous-espaces de dimension $n = n(h)$ ($n(h) \to \infty$ pour $h \to 0$) qui approchent U au sens qu'on trouve pour tout $v(x) \in U$ et tout $\Delta > 0$ un h_Δ tel qu'il existe pour tous les $h < h_\Delta$ un élément $v^h(x) \in U^h$ qui approche $v(x)$ à Δ près, i.e. tout $h < h_\Delta$ vérifie l'inégalité

$$\inf_{\varphi \in U^h} \|v - \varphi\| \leqslant \Delta. \tag{8.48}$$

On remplace (8.3) par le problème approché

$$J(u^h) = \inf_{v \in U_d^h} J(v), \tag{8.49}$$

$U_d^h = U_d \cap U^h$ étant un ensemble convexe fermé (l'intersection de deux ensembles convexes fermés est un convexe fermé). Le problème (8.49) admet évidemment une solution unique (voir point 2.8.1) qui converge pour $h \to 0$ vers la solution du problème initial (8.3) par suite de l'hypothèse de U^h complète dans U.

Supposons que les fonctions $\psi_k^h(x)$, $k = 1, \ldots, n$, forment une base de l'espace U^h et qu'une fonction $v(x)$ respecte la relation

$$v(x) = \sum_{k=1}^{n} \alpha_k^h \psi_k^h(x). \tag{8.50}$$

Sous ces hypothèses, on a

$$J(v) \equiv \Phi(\alpha) = (A\alpha, \alpha) - 2(g, \alpha), \tag{8.51}$$

où A est une matrice définie positive symétrique d'ordre n formée de

$$a_{ij} = \pi(\psi_i^h, \psi_j^h), \tag{8.52}$$

g et α sont des vecteurs de dimension n ayant pour composantes

$$\begin{aligned} g_i &= L(\psi_i^h), \\ \alpha_i &= \alpha_i^h \end{aligned} \tag{8.53}$$

et (,) désigne le produit scalaire ordinaire dans l'espace engendré par des vecteurs réels de dimension n. Le problème (8.49) se récrit

$$\Phi(\beta) = \inf_{\alpha \in V_d} \Phi(\alpha), \tag{8.54}$$

avec V_d défini comme suit: $\alpha \in V_d$ si et seulement si la fonction

$$v(x) = \sum_{k=1}^{n} \alpha_k \psi_k^h(x) \tag{8.55}$$

appartient à U_d^h. V_d est évidemment un ensemble convexe fermé de l'espace des vecteurs réels n-dimensionnels et le problème (8.55) admet une solution unique β, la fonction

$$u^h(x) = \sum_{k=1}^{n} \beta_k \psi_k^h(x)$$

étant la solution de (8.49).

On connaît de nombreux algorithmes de résolution pour le problème (8.54), dont les plus utilisés sont diverses variantes de la méthode de descente, les méthodes du gradient, les directions admissibles, les variations locales. Pour plus de détail voir les monographies de Pchénitchny et Daniline [III], Tchernoousko, Banitchouk [III; 16] et Polak [III].

Décrivons une version de la méthode de relaxation.

Soit $\alpha^{(s)} = (\alpha_1^{(s)}, \ldots, \alpha_n^{(s)}) \in V_d$ la s-ième approximation de la solution β de (8.54). On en obtient la $(s+1)$-ième après n pas fractionnaires par les formules

$$\alpha^{s+\frac{i}{n}} = \alpha^{s+\frac{i-1}{n}} - q_i^s r_i^s e_i, \qquad i = 1, \ldots, n, \tag{8.56}$$

e_i étant un vecteur de composantes $e_{ij} = \delta_{ij}$, $j = 1, \ldots, n$ (δ_{ij} est le symbole de Kronecker),

$$r_i^s = \sum_{j=1}^{n} a_{ij} \alpha_j^{s+\frac{i-1}{n}} - g_i = \left.\frac{\partial \Phi(\alpha)}{\partial \alpha_i}\right|_{\alpha = \alpha^{s+\frac{i-1}{n}}}, \tag{8.57}$$

et q_i^s la solution du problème d'extrémum de dimension un

$$\Phi\left(\alpha^{s+\frac{i}{n}}\right) = \inf_q \Phi\left(\alpha^{s+\frac{i-1}{n}} - q r_i^s e_i\right), \qquad \alpha^{s+\frac{i}{n}} \in V_d. \tag{8.58}$$

Si V_d coïncide avec l'espace tout entier, ce procédé de descente séquentielle devient la méthode connue de Gauss-Seidel pour les systèmes d'équations linéaires $A\alpha = g$ à matrice A définie positive symétrique.

Prenons par exemple le problème (8.19), où U_d se définit par (8.37) sous l'hypothèse supplémentaire de $F_1(x)$ concave et $F_2(x)$ convexe, i.e. n'importe quels x', x'' de [0, 1] et $\Theta \in [0, 1]$ vérifient les inégalités

$$\begin{aligned} F_1((1-\Theta)x' + \Theta x'') &\geqslant (1-\Theta) F_1(x') + \Theta F_1(x''), \\ F_2((1-\Theta)x' + \Theta x'') &\leqslant (1-\Theta) F_2(x') + \Theta F_2(x''). \end{aligned} \tag{8.59}$$

On associe à [0, 1] pour $n \geqslant 1$ donné le réseau régulier

$$\Omega_h = \left\{x_k \mid x_k = k \times h, \quad k = 0, 1, \ldots, N+1, \ h = \frac{1}{N+1}\right\}, \tag{8.60}$$

et on introduit par (3.14) un système de fonctions de base $\psi_k(x)$, $k = 1, \ldots, n$. L'espace U^h est alors défini comme étant l'enveloppe linéaire du système $\{\psi_k(x)\}_{k=1}^n$. Autrement dit, U^h est l'espace des prolongements linéaires par morceaux des fonctions discrètes définies aux nœuds de Ω_h et nulles aux points $x_0 = 0$ et $x_{N+1} = 1$.

Etant données les hypothèses faites sur $F_1(x)$, $F_2(x)$ et U^h, l'ensemble $U_d^h = U_d \cap U^h$ est celui des fonctions $v(x) \in U^h$ telles que

$$F_1(x_i) \leqslant v(x_i) \leqslant F_2(x_i), \qquad i = 1, \ldots, n,$$

et V_d constitue un parallélépipède rectangle de dimension n dont les coordonnées $\alpha = (\alpha_1, \ldots, \alpha_n)$ satisfont aux inégalités

$$F_1(x_i) \leqslant \alpha_i \leqslant F_2(x_i), \qquad i = 1, \ldots, n.$$

Si l'on se passe des propriétés supposées de $F_1(x)$ et $F_2(x)$ d'être respectivement concave et convexe, l'appartenance $v(x) \in U_d^h$ (et $\alpha \in V_d$) risque d'avoir lieu sous des conditions plus sévères.

Illustrons la méthode des variations locales sur un problème variationnel unidimensionnel fort simple. Soit la fonctionnelle variationnelle

$$J(\varphi) = \int_0^1 \pi(x, \varphi, \varphi_x)\, dx. \tag{8.61}$$

On demande la fonction $\varphi(x)$ qui réalise le minimum de $J(\varphi)$ et vérifie les contraintes

$$F_1(x) \leqslant \varphi(x) \leqslant F_2(x). \tag{8.62}$$

On observe que les contraintes ne s'imposent nécessairement pas sur [0, 1] tout entier, si bien qu'on est en présence d'une contrainte naturelle formelle : ou bien

$$-\infty < \varphi,$$

ou bien

$$\varphi < \infty.$$

Cette interprétation du problème autorise à assimiler les conditions aux limites pour $\varphi(x)$, $x = 0$, $x = 1$, à des contraintes.

Voici une mise en œuvre simple de la méthode en question. On partage [0, 1] en n parties par les points équidistants $x_i = i\Delta x$, $\Delta x = \frac{1}{n}$, $i = 0, 1, \ldots, n$, et on représente la fonctionnelle (8.61) par la somme de fonctionnelles partielles

$$J = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \pi(x, \varphi, \varphi_x)\, dx, \tag{8.63}$$

chaque intégrale étant approchée par la formule des rectangles

$$\int_{x_i}^{x_{i+1}} \pi(x, \varphi, \varphi_x)\, dx \approx \Delta x \pi(x_{i+1/2}, \varphi_{i+1/2}, \varphi'_{i+1/2}),$$

où

$$\varphi_{i+1/2} = \frac{\varphi_i + \varphi_{i+1}}{2}, \quad \varphi'_{i+1/2} = \frac{\varphi_{i+1} - \varphi_i}{\Delta x}, \quad x_{i+1/2} = x_i + \frac{\Delta x}{2}.$$

Ainsi, seules les valeurs de la fonction cherchée aux points φ_i et φ_{i+1} figureront dans les fonctionnelles partielles. On introduit la notation

$$V_i(\varphi_i, \varphi_{i+1}) = \Delta x \pi(x_{i+1/2}, \varphi_{i+1/2}, \varphi'_{i+1/2}).$$

On a

$$J(\varphi) \approx V = \sum_{i=0}^{n-1} V_i(\varphi_i, \varphi_{i+1}). \tag{8.64}$$

Ainsi, nous sommes amené au problème de chercher un ensemble de nombres $\{\varphi_i\}$ qui minimise la fonctionnelle V avec les contraintes données

$$F_1(x_i) \leqslant \varphi_i \leqslant F_2(x_i). \tag{8.65}$$

On procède par approximations successives. Supposons connues k premières approximations φ_i^k. On cherche φ_i^{k+1} en introduisant un nombre h et en associant à chaque x_i le triplet $\varphi_i^k - h$, φ_i^k, $\varphi_i^k + h$. La quantité φ_i^k vérifie les contraintes par hypothèse. Vérifions s'il en est de même de $\varphi_i^k - h$, $\varphi_i^k + h$. Si une contrainte se trouve violée, on rejette le nombre concerné et on opère soit sur $\varphi_i^k - h$, φ_i^k, soit sur φ_i^k, $\varphi_i^k + h$.

Comme dans la méthode de Gauss-Seidel, on suppose connues les valeurs de la $(k+1)$-ième approximation pour tous les indices inférieurs à i, i.e. pour φ_0^{k+1}, φ_1^{k+1}, . . ., φ_{i-1}^{k+1}. Il est naturel qu'on utilise cette information pour accélérer la convergence. Aussi on introduit, en plus des V_i, les fonctionnelles partielles

$$\begin{aligned} G_i &= V_{i-1}(\varphi_{i-1}^{k+1},\ \varphi_i^k) + V_i(\varphi_i^k,\ \varphi_{i+1}^k), \\ U_i^+ &= V_{i-1}(\varphi_{i-1}^{k+1},\ \varphi_i^k + h) + V_i(\varphi_i^k + h,\ \varphi_{i+1}^k), \\ U_i^- &= V_{i-1}(\varphi_{i-1}^{k+1},\ \varphi_i^k - h) + V_i(\varphi_i^k - h,\ \varphi_{i-1}^k). \end{aligned} \tag{8.66}$$

Ces expressions tiennent compte de deux termes de la somme (8.64) qui dépendent de φ_i et correspondent aux trois variantes possibles $\varphi_i^k - h$, φ_i^k, $\varphi_i^k + h$.

On définit celui de ces nombres qui vérifie les contraintes données et réalise le minimum de l'une des trois fonctionnelles (8.66). On accepte ce nombre en qualité de la valeur approchée suivante φ_i^{k+1}. On cherche les valeurs de φ_i^{k+1} pour tous les $i = 0, \ldots, n$, et ainsi de suite.

Une fois la solution approchée obtenue pour Δx et h donnés avec une précision désirée, le processus est repris avec une nouvelle valeur de h, disons h diminué de moitié, et ainsi de suite. On choisit naturellement au départ ladite solution approchée. On réduit ensuite Δx et on affine le résultat.

Le cas considéré a été très simple du point de vue contraintes: la solution φ même a été astreinte à certaines conditions. Dans le cas plus général où la contrainte est définie par la relation

$$B(x,\ \varphi,\ \varphi_x) \leqslant 0,$$

on utilise pour tout x_i telle ou telle approximation et on teste toutes les fois cette condition approchée.

La convergence et l'exactitude de la méthode des variations locales sont discutées en détail par Tchernoousko et Banitchouk [III ; 16].

La méthode des variations locales fonctionne également pour des problèmes variationnels pluridimensionnels. On passe en principe par toutes les étapes de l'algorithme décrit. En dimension deux, on a, au lieu de $\varphi_i - h$, φ_i, $\varphi_i + h$ à identifier moyennant les fonction-

nelles et les contraintes, les nombres $\varphi_{ij} - h$, φ_{ij}, $\varphi_{ij}+h$, et la fonctionnelle partielle U_{ij} est la somme de quatre termes

$$U_{ij}=V_{ij}+V_{i-1,\,j}+V_{i,\,j-1}+V_{i-1,\,j-1}$$

en φ^h_{ij}. On définit de même U^+_{ij} et U^-_{ij}, avec $\varphi_{ij} = \varphi_{ij}+h$ et $\varphi_{ij} = \varphi_{ij} - h$ respectivement. On a un algorithme analogue pour des problèmes de dimension supérieure.

Chose importante, la méthode des variations locales n'étant pas liée au choix de la base joue pour les domaines de forme quelconque. D'autre part, elle peut conduire à un minimum local. Quant à la convergence vers le minimum local, elle est une propriété intrinsèque de l'algorithme.

CHAPITRE 3

INTERPOLATION DES FONCTIONS DISCRÈTES

Le problème d'interpoler des quantités définies sur un ensemble discret de points sur tout le domaine de définition d'une fonction de l'argument continu est intimement lié à la génération de schémas variationnels aux différences et à la représentation continue des solutions de problèmes aux différences. En effet, on construit de règle les équations aux différences en discrétisant l'opérateur et la solution du problème par des méthodes de projection convenables. La solution du problème aux différences constitue en général une approximation de la solution du problème initial sur un ensemble ponctuel discret. Supposons résolu le problème aux différences, si bien que nous possédons l'information sur la solution approchée. Il s'agit maintenant d'interpoler ces données sur tout le domaine de définition de la solution du problème primitif tout en respectant certaines conditions, savoir : si la solution de l'équation aux différences est obtenue avec une précision déterminée, il faut qu'il y ait concordance entre l'ordre d'interpolation de données et l'ordre d'approximation de cette équation et que le premier soit au moins égal au second. En présence d'une information auxiliaire sur les erreurs affectant la solution approchée, on interpole non pas par rapport aux données exactes, mais en prenant en considération les erreurs éventuelles sur les valeurs aux nœuds. Il se peut que l'information à priori sur la régularité de la solution permette alors d'améliorer l'approximation obtenue par une technique aux différences. D'autre part, l'interpolation de données constitue un problème à part.

Les algorithmes d'interpolation de fonctions définies exactement sur un ensemble ponctuel discret utilisent en général les polynômes d'interpolation de Lagrange et on suppose à priori la fonction à interpoler $\varphi(x)$ dérivable jusqu'à un certain ordre.

Un problème voisin de l'interpolation est celui où $\varphi(x)$ donnée est affectée aux nœuds x_k d'une erreur dont on connaît à priori la borne supérieure pour chaque point. Il consiste à construire une courbe qui l'approche au mieux dans un sens et on l'aborde en général par les moindres carrés.

La dernière acquisition de la théorie de l'interpolation porte le nom de méthodes d'*interpolation de fonctions splines*. On appelle communément *fonction spline* une fonction polynomiale par morceaux définie dans un domaine D, i.e. c'est une fonction telle qu'on peut partager D en des domaines partiels de façon qu'elle soit dans chacun un polynôme de degré m.

Cette fonction est en général continue dans D ainsi que ses dérivées jusqu'à l'ordre $m - 1$ et possède une dérivée m-ième de carré intégrable. Les problèmes de la technique utilisent essentiellement les fonctions splines du troisième degré.

L'interpolation à l'aide de fonctions splines est étudiée avec force détails par Schoenberg [XIV], Holladay (XIV], Ahlberg, Nilson, Walsh [III], Zavialov [XIV; 5]. On trouve diverses généralisations de la théorie des splines dans Birkhoff, Schultz, Varga [V], Anselone, Laurent [XIV], Morozov [XIV; 8], Vassilenko [XIV; 2].

3.1. Interpolation des fonctions d'une variable

3.1.1. Interpolation des fonctions d'une variable par les fonctions splines du troisième degré

Soit, sur le segment $[a, b]$ de l'axe réel x, le réseau $a = x_0 < x_1 < \ldots < x_n = b$ et soit $f(x)$ une fonction définie sur $[a, b]$ à valeurs $\{f_k\}_{k=0}^n$ données aux nœuds du réseau. Le problème d'interpolation cubique par morceaux s'énonce de façon suivante : on cherche sur $[a, b]$ une fonction $g(x)$ ayant les propriétés :

1) $g(x)$ est de classe $C^{(2)}(a, b)$, i.e. elle est continue ainsi que ses dérivées de deux premiers ordres ;

2) sur chaque $[x_{k-1}, x_k]$, elle est un polynôme du troisième degré de la forme

$$g(x) \equiv g_k(x) = \sum_{l=0}^{3} a_l^{(k)} (x_k - x)^l, \qquad k = 1, \ldots, n; \tag{1.1}$$

3) on a aux points $\{x_k\}_{k=0}^n$ du réseau

$$g(x_k) = f_k, \qquad k = 0, 1, \ldots, n; \tag{1.2}$$

4) $g''(x)$ vérifie les conditions aux limites

$$g''(a) = g''(b) = 0. \tag{1.3}$$

Nous ne pouvons apprécier les mérites de cette interpolation qu'après avoir établi la propriété extrémale naturelle de $g(x)$ concrète.

On montre l'unicité pour le problème de recherche de $g(x)$ et on utilise à cet effet les conditions 1)-4).

Etant données la continuité et la linéarité sur chaque $[x_{i-1}, x_i]$, $i = 1, \ldots, n$, de la dérivée seconde de $g(x)$, on écrit pour $x_{i-1} \leqslant x \leqslant x_i$

$$g''(x) = m_{i-1}\frac{x_i - x}{h_i} + m_i\frac{x - x_{i-1}}{h_i}, \tag{1.4}$$

où $h_i = x_i - x_{i-1}$, $m_k = g''(x_k)$. On intègre deux fois les deux membres de (1.4), il vient

$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + A_i\frac{x_i - x}{h_i} + B_i\frac{x - x_{i-1}}{h_i}, \tag{1.5}$$

avec A_i et B_i des constantes d'intégration qui s'obtiennent de suite à partir de la condition $g(x_{i-1}) = f_{i-1}$, $g(x_i) = f_i$. Portons $x = x_i$ et $x = x_{i-1}$ dans (1.5):

$$m_i\frac{h_i^2}{6} + B_i = f_i,$$

$$m_{i-1}\frac{h_i^2}{6} + A_i = f_{i-1}.$$

Finalement,

$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(f_i - \frac{m_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}, \tag{1.6}$$

$$g'(x) = -m_{i-1}\frac{(x_i - x)^2}{2h_i} + m_i\frac{(x - x_{i-1})^2}{2h_i} + \frac{f_i - f_{i-1}}{h_i} - \frac{m_i - m_{i-1}}{6}h_i. \tag{1.7}$$

La dernière égalité fournit les limites à droite et à gauche aux points $x_1, x_2, \ldots, x_{n-1}$:

$$g'(x_i - 0) = \frac{h_i}{6}m_{i-1} + \frac{h_i}{3}m_i + \frac{f_i - f_{i-1}}{h_i},$$

$$g'(x_i + 0) = -\frac{h_{i+1}}{3}m_i - \frac{h_{i+1}}{6}m_{i+1} + \frac{f_{i+1} - f_i}{h_{i+1}}.$$

Selon 1), $g''(x)$ et $g'(x)$ sont continues sur $[a, b]$. La continuité de $g'(x)$ en $x_1, x_2, \ldots, x_{n-1}$ donne, compte tenu de (1.7), $n-1$ équations

$$\frac{h_i}{6}m_{i-1} + \frac{h_i + h_{i+1}}{3}m_i + \frac{h_{i+1}}{6}m_{i+1} = \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}. \tag{1.8}$$

Complétons ce système par les égalités $m_0 = m_n = 0$ (condition(1. 3)), il vient le système algébrique linéaire pour déterminer les inconnues $m_1, m_2, \ldots, m_{n-1}$:

$$Am = Hf, \tag{1.9}$$

A étant une matrice carrée telle que

$$A = \begin{Vmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \dots & 0 & 0 \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \dots & 0 & 0 \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{h_{n-1}}{6} & \frac{h_n+h_{n-1}}{3} \end{Vmatrix} \tag{1.10}$$

et H, m et f respectivement une matrice rectangulaire et deux vecteurs de la forme

$$m = \begin{Vmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n-1} \end{Vmatrix}, \qquad f = \begin{Vmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{Vmatrix},$$

$$H = \begin{Vmatrix} \frac{1}{h_1} & \left(-\frac{1}{h_1}-\frac{1}{h_2}\right) & \frac{1}{h_2} & \dots & 0 & 0 \\ 0 & \frac{1}{h_2} & \left(-\frac{1}{h_2}-\frac{1}{h_3}\right) & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \left(-\frac{1}{h_{n-1}}-\frac{1}{h_n}\right) & \frac{1}{h_n} \end{Vmatrix}. \tag{1.11}$$

A est symétrique à diagonale dominante stricte. Aux termes du théorème de Gershgorin, elle est définie positive et visiblement non singulière, si bien que le système (1.9) définit les coefficients m_1, m_2, . . ., m_{n-1} de façon unique. La fonction spline $g(x)$ est donc reconstituée de façon unique par les formules (1.6), d'où l'unicité pour le problème proposé.

Les fonctions splines du troisième degré jouissent d'une propriété remarquable qui garantit l'efficacité de l'interpolation à l'aide de splines. Considérons sur le segment $[a, b]$ la classe $W_2^2[a, b]$ de fonctions à dérivées secondes de carré sommable. On demande la fonction d'interpolation $u \in W_2^2[a, b]$,

$$u(x_k) = f_k, \quad k = 0, 1, \dots, n, \tag{1.12}$$

qui minimise la fonctionnelle

$$\Phi(u) = \int_a^b [u''(x)]^2 \, dx \tag{1.13}$$

sur la classe $W_2^2[a, b]$. Je dis que le minimum de (1.13) est réalisé par $g(x)$ que nous venons de construire. En effet, soit

$$\Phi(u-g)=\int_a^b [u''-g'']^2\,dx. \tag{1.14}$$

Intégrons par parties et utilisons les propriétés de g et $u \in W_2^2$, il vient

$$\Phi(u-g)=\Phi(u)-\Phi(g)-2\left[(u'-g')\,g''\Big|_{x=a}^{x=b}-\int_a^b (u'-g')\,g'''\,dx\right]=$$

$$=\Phi(u)-\Phi(g)-2\sum_{k=1}^{n}\int_{x_{k-1}}^{x_k}(u'-g')\,g'''\,dx.$$

Mais $g'''=c_k=\text{const}$ sur $[x_{k-1}, x_k]$, donc

$$\Phi(u-g)=\Phi(u)-\Phi(g)-2\sum_{k=1}^{n} c_k\,(u-g)\Big|_{x=x_{k-1}}^{x=x_k}=\Phi(u)-\Phi(g),$$

et on a

$$\Phi(g)=\Phi(u)-\Phi(u-g)\leqslant\Phi(u) \tag{1.15}$$

pour toute $u \in W_2^2$, $u(x_k)=f_k$, $k=0, 1, \ldots, n$. Ainsi, la fonction spline $g(x)$ réalise le minimum de la fonctionnelle (1.13). On montre l'unicité de ce point de minimum et on donne, moyennant (1.12), (1.13), la définition équivalente de la fonction spline du troisième degré: c'est une fonction de classe $W_2^2[a, b]$ qui prend des valeurs données aux nœuds du réseau et minimise la fonctionnelle (1.13). Cette propriété découverte par Holladay [XIV] paraît très prometteuse. En effet, on interprète désormais la fonctionnelle $\Phi(u)$ comme un analogue de l'énergie potentielle d'une tige élastique fixée aux points du plan (x_k, f_k) dont le minimum est réalisé par des fonctions splines du troisième degré.

Les êtres mathématiques en question possèdent de bonnes propriétés d'approximation. Si la fonction à interpoler f est de classe $C^k[a, b]$, $k=0, 1, 2, 3, 4$, alors la fonction d'erreur $\varphi(x)=f(x)-g(x)$ vérifie les inégalités

$$\max_{a\leqslant x\leqslant b}|\varphi^{(p)}(x)|\leqslant Ch^{k-p}, \qquad k=0, 1, 2, 3, 4, \qquad k\geqslant p, \tag{1.16}$$

où la constante non négative C ne dépend pas du réseau,

$$h=\max_{1\leqslant k\leqslant n}|x_k-x_{k-1}|.$$

L'estimation (1.16) est rude. On en trouve de meilleures dans Varga [I].

3.1.2. Interpolation cubique par morceaux avec lissage

Nous allons considérer le problème de prolonger régulièrement une fonction définie sur le réseau

$$a = x_0 < x_1 < \ldots < x_n = b,$$

les valeurs aux nœuds $\tilde{f}_k$ étant affectées d'une erreur. Cette fois, on n'a que faire d'un interpolant qui se confond aux nœuds avec les valeurs données. On cherche une fonction dont la courbe représentative soit plus lisse et passe plus près de ces points que celle de la fonction d'interpolation. Ces fonctions sont dites *de lissage* (ou *d'ajustement*).

Imposons à $g(x)$ lissante cherchée de minimiser sur la classe $W_2^2[a, b]$ la fonctionnelle

$$\Phi_1(u) = \int_a^b [u'']^2\,dx + \sum_{k=0}^{n} p_k [u(x_k) - \tilde{f}_k]^2, \tag{1.17}$$

p_k étant des nombres positifs. La fonctionnelle $\Phi_1(u)$ exprime deux conditions : *primo*, la courbe passe près des points donnés ; *secundo*, la courbe est lisse au maximum. La fonction d'ajustement passe d'autant plus près des points donnés que les coefficients de poids p_k sont élevés.

On montre que le problème variationnel (1.17) admet pour solution une fonction spline du troisième degré, i.e. une fonction vérifiant les conditions 1), 2) et 4) du point précédent. Soit $u_0 \in W_2^2[a, b]$ la solution cherchée. On forme une fonction spline $g(x)$ telle que $g(x_k) = u_0(x_k)$, $k = 0, 1, \ldots, n$. Le second terme de (1.17) est le même pour $g(x)$ et $u_0(x)$, si bien que

$$\int_a^b [u_0'']^2\,dx \leqslant \int_a^b [g'']^2\,dx. \tag{1.18}$$

Or, on a montré au point précédent que $g(x)$ réalise, dans l'interpolation de $u_0(x)$, le seul minimum de $\int_a^b [u'']\,dx$. Donc $u_0 \equiv g$.

Ainsi, il suffit de chercher le minimum de $\Phi_1(u)$ dans la classe de fonctions splines du troisième degré. Comme une fonction spline du troisième degré se définit de façon unique par l'ensemble de ses valeurs $\{\mu_k\}_{k=0}^n$ aux nœuds $\{x_k\}_{k=0}^n$, la minimisation de $\Phi_1(u)$ se ramène

à la recherche du minimum d'une fonction des variables $\mu_0, \mu_1, \ldots, \mu_n$, $g''(x)$ étant linéaire par morceaux,

$$g''(x) = m_{k-1}\frac{x_k - x}{h_k} + m_k\frac{x - x_{k-1}}{h_k} \tag{1.19}$$

pour $x \in [x_{k-1}, x_k]$, $m_k = g''(x_k)$, $k = 1, \ldots, n-1$, $m_0 = m_n = 0$. Donc

$$\Phi_1(g) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \left[m_{k-1}\frac{x_k - x}{h_k} + m_k\frac{x - x_{k-1}}{h_k}\right]^2 dx + \sum_{k=0}^{n} p_k(\mu_k - \tilde{f}_k)^2. \tag{1.20}$$

Effectuons l'intégration, il vient

$$\sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \left[m_{k-1}\frac{x_k - x}{h_k} + m_k\frac{x - x_{k-1}}{h_k}\right]^2 dx =$$

$$= \sum_{k=1}^{n} m_k\left[m_{k-1}\frac{h_k}{6} + \frac{h_k + h_{k+1}}{3}m_k + \frac{h_{k+1}}{6}m_{k+1}\right] = (Am, m), \tag{1.21}$$

avec A la matrice (1.10). Ainsi

$$\Phi_1(g) = (Am, m) + \sum_{k=0}^{n} p_k(\mu_k - \tilde{f}_k)^2. \tag{1.22}$$

Etant donné (1.9), m dépend linéairement des $\mu = (\mu_0, \mu_1, \ldots, \mu_n)$, si bien que $\Phi_1(g)$ est une forme définie positive de μ. Cette forme ne peut avoir d'autres points extrémaux qu'un minimum, et ce sous la condition nécessaire

$$\frac{\partial \Phi_1}{\partial \mu_s} \equiv \frac{\partial}{\partial \mu_s}(Am, m) + 2p_s(\mu_s - \tilde{f}_s) = 0, \qquad s = 0, 1, \ldots, n.$$

Mais la matrice A ne dépend pas de μ. Aussi on a, en vertu de (1.9),

$$\frac{\partial}{\partial \mu_s}(Am, m) = 2\left(\frac{\partial (Am)}{\partial \mu_s}, m\right) = 2\left(\frac{\partial (H\mu)}{\partial \mu_s}, m\right) =$$

$$= 2\left(\frac{\partial \mu}{\partial \mu_s}, H^*m\right) = 2(H^*m)_s,$$

avec H définie par (1.11). Il en résulte la condition de minimum sous forme vectorielle

$$H^*m + P\mu = P\tilde{f}, \tag{1.23}$$

où le vecteur $\tilde{f} = (\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_n)$ et la matrice P est diagonale :

$$P = \begin{Vmatrix} p_0 & 0 & \ldots & 0 \\ 0 & p_1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & p_n \end{Vmatrix}. \tag{1.24}$$

Prémultiplions (1.23) par HP^{-1} :

$$HP^{-1}H^*m + H\mu = H\tilde{f},$$

ou encore, vu (1.9),

$$(A + HP^{-1}H^*)\, m = H\tilde{f}. \tag{1.25}$$

La matrice de ce système est 5-diagonale, symétrique et définie positive. On résout (1.25) par exemple par la méthode d'élimination de Gauss. Une fois le vecteur m défini, on trouve le vecteur valeurs discrètes de la spline d'ajustement par la formule

$$\mu = \tilde{f} - P^{-1}H^*m \tag{1.26}$$

qui résulte de suite de (1.23), et on calcule $g(x)$ par (1.6).

3.1.3. Prolongements réguliers

Riabenki [XIV ; 10] a proposé un procédé de prolongement régulier de fonctions discrètes qui diffère légèrement des méthodes de fonctions splines, mais qui donne lieu à des algorithmes non moins efficaces. Voici, très brièvement, le schéma de formation d'une fonction d'interpolation de classe de régularité C^p arbitraire.

Soit $x_1 < x_2 < \ldots < x_{n-1} < x_n$ un réseau fixe et f une fonction dont on connaît les valeurs nodales $f_1, f_2, \ldots, f_n$. On suppose n suffisamment important et on fixe un entier $p \ll n$. Construisons sur (x_1, x_2) une fonction d'interpolation de classe C^p. On désigne par $P_0(x)$ le polynôme de Lagrange de degré p au plus qui prend des valeurs données aux nœuds $x_1, x_2, \ldots, x_{p+1}$ et par $P_1(x)$ le polynôme de Lagrange de degré p au plus qui prend des valeurs données aux nœuds $x_2, x_3, \ldots, x_{p+2}$. On forme un polynôme $Q_1(x)$ de degré $2p + 1$ au plus tel qu'on ait

$$\left.\frac{d^kQ_1}{dx^k}\right|_{x=x_1} = \left.\frac{d^kP_0}{dx^k}\right|_{x=x_1}, \quad \left.\frac{d^kQ_1}{dx^k}\right|_{x=x_2} = \left.\frac{d^kP_1}{dx^k}\right|_{x=x_2}, \tag{1.27}$$

$$k = 0, 1, \ldots, p.$$

Ce polynôme existe évidemment et les conditions (1.27) le définissent pe façon univoque. Admettons que la fonction d'interpolation $g(x)$ relative à (x_1, x_2) est égale au polynôme $Q_1(x)$.

Prenons comme exemple le cas $p = 1$. Le polynôme $Q_1(x)$ du troisième degré associé à (x_1, x_2) s'écrit

$$Q_1(x) = a_0 + a_1(x - x_1) + a_2(x - x_1)^2 + a_3(x - x_1)^3$$

et ses coefficients sont calculés par les formules

$$\begin{aligned} a_0 &= f_1, \\ a_1 &= \frac{f_2 - f_1}{x_2 - x_1}, \\ a_2 &= -\frac{1}{x_2 - x_1}\left(\frac{f_3 - f_2}{x_3 - x_2} - \frac{f_2 - f_1}{x_2 - x_1}\right), \\ a_3 &= \frac{1}{(x_2 - x_1)^2}\left(\frac{f_3 - f_2}{x_3 - x_2} - \frac{f_2 - f_1}{x_2 - x_1}\right). \end{aligned} \tag{1.28}$$

On procède de même pour $g(x)$ et le segment (x_2, x_3), et ainsi de suite jusqu'à l'intervalle (x_{n-p-1}, x_{n-p}). L'interpolant $g(x)$ ainsi construit est une fonction morceaux de polynômes de classe C^p, les polynômes étant de degré $2p + 1$ sur les intervalles du réseau. Avec $p = 2$ (ce qui correspond à la régularité des fonctions splines du troisième degré), les polynômes obtenus sont de degré 5, mais on n'a pas en revanche à résoudre le système algébrique linéaire.

La construction de fonctions d'interpolation de plusieurs variables diffère peu du cas unidimensionnel (elle s'y ramène d'ailleurs). Ce sont justement les problèmes pluridimensionnels qui ont intéressé Riabenki.

Les prolongements réguliers jouissent de bonnes propriétés d'approximation. En interpolant une fonction $f \in C^q$ de plusieurs variables par $g(x)$ de classe C^p, $p \geqslant q$, la fonction et ses dérivées vérifient les inégalités

$$\| D^k f - D^k g \|_C \leqslant C(p)\, h^{q-|k|} \sup_x \max_{|\alpha|=q} | D^\alpha f(x) |, \tag{1.29}$$

où le multi-indice $k = (k_1, k_2, \ldots, k_m)$ est tel que $|k| = \sum_{i=1}^{m} k_i < q$, $D^k = \dfrac{\partial^{k_1 + \ldots + k_m}}{\partial x^{k_1} \ldots \partial x^{k_m}}$ et $C(p)$ dépend de p seul (elle est indépendante du pas h et de la fonction f).

Il y a lieu de signaler une autre propriété des prolongements réguliers. Toute fonction de base, i.e. une fonction égale à 1 en un nœud et à 0 partout ailleurs, est à support borné pour tout p et le « rayon » du support ne dépasse pas p. Mais si l'on accepte pour fonction de base une spline du troisième degré, il s'avère qu'elle n'est en général pas à support borné. On conçoit que les prolongements réguliers de Riabenki sont tout indiqués dans les problèmes variationnels abordés par les éléments finis.

3.2. Interpolation à deux et à plusieurs variables

L'interpolation en dimension deux à l'aide de fonctions splines bicubiques (i.e. elles sont du troisième degré par rapport à chacune des variables) a été étudiée entre autres par Ahlberg, Nilson, Walsh [III] et Zavialov [XIV; 5]. Considérons le problème modèle suivant.

Soit $D = \{x, y : a \leqslant x \leqslant b, c \leqslant y \leqslant d\}$ un domaine rectangulaire. Superposons à D le réseau $D_h = \{x_k, y_i : a = x_0 < x_1 < \ldots < x_n = b; c = y_0 < y_1 < \ldots < y_m = d\}$. Avec les hypothèses faites, le problème d'interpoler une fonction $f(x, y)$ définie aux points de D_h par des splines bicubiques consiste à chercher la fonction $g(x, y)$ vérifiant les conditions

1) $$g(x, y) \in C^2(D); \tag{2.1}$$

2) dans chaque cellule $g(x, y)$ est un polynôme bicubique de la forme

$$g(x, y) = g_{k,l}(x, y) = \sum_{i,j=0}^{3} a_{i,j}^{k,l} (x_k - x)^i (y_l - y)^j; \tag{2.2}$$

3) sur D_h, $g(x, y)$ prend les valeurs données

$$g(x_k, y_l) = f_{kl}, \qquad k = 0, \ldots, n; \; l = 0, \ldots, m; \tag{2.3}$$

4) $g(x, y)$ remplit la condition

$$\frac{\partial^2 g}{\partial \nu^2}\Big|_\Gamma = 0, \tag{2.4}$$

ν étant la normale extérieure à la frontière Γ de D.

Le problème ne diffère en principe pas du cas à une variable. On rappelle qu'on détermine une fonction spline à une variable en tout point par les formules simples (1.6) connaissant ses valeurs aux nœuds, ainsi que les valeurs de ses dérivées secondes. On trouve les dérivées secondes en résolvant un système algébrique linéaire à matrice 3-diagonale. Voyons ce que deviennent ces calculs préliminaires en dimension deux.

Considérons les problèmes unidimensionnels d'interpoler par les fonctions splines du troisième degré basées sur les lignes $y = y_j$, $j = 0, 1, \ldots, m$, ce qui exige la résolution de $m + 1$ systèmes algébriques linéaires du type (1.9). On trouve les valeurs de $g_{xx}(x, y)$ aux nœuds de D_h. On procède de même pour $n + 1$ problèmes associés aux lignes $x = x_i$, $i = 0, 1, \ldots, n$, ce qui donne les valeurs de $g_{yy}(x, y)$ sur D_h. On se propose de calculer $g(x, y)$ en un point

(x, y). Soit $x_{i-1} \leqslant x \leqslant x_i$, $y_{j-1} \leqslant y \leqslant y_j$. On définit facilement $g(x, y)$ aux points (x_i, y), (x_{i-1}, y) par les formules analogues à (1.6)

$$g(x_{i-1}, y) = N_{i-1\,j-1} = \frac{(y_j - y)^3}{6\tau_j} + N_{i-1\,j}\frac{(y - y_{j-1})^3}{6\tau_j} +$$
$$+ \left(f_{i-1\,j-1} - \frac{N_{i-1\,j-1}\tau_j^2}{6}\right)\frac{y_j - y}{\tau_j} + \left(f_{i-1\,j} - \frac{N_{i-1\,j}\tau_j^2}{6}\right)\frac{y - y_{j-1}}{\tau_j}, \quad (2.5)$$

$$g(x_i, y) = N_{i\,j-1}\frac{(y_j - y)^3}{6\tau_j} + N_{ij}\frac{(y - y_{j-1})^3}{6\tau_j} +$$
$$+ \left(f_{i\,j-1} - \frac{N_{i\,j-1}\tau_j^2}{6}\right)\frac{y_j - y}{\tau_j} + \left(f_{ij} - \frac{N_{ij}\tau_j^2}{6}\right)\frac{y - y_{j-1}}{\tau_j}. \quad (2.6)$$

On a introduit les notations

$$N_{ij} = g_{yy}(x_i, y_i), \qquad M_{ij} = g_{xx}(x_i, y_j),$$
$$h_i = x_i - x_{i-1}, \qquad \tau_j = y_j - y_{j-1}.$$

Si l'on connaissait les valeurs de $g_{xx}(x_{i-1}, y)$ et $g_{xx}(x_i, y)$, on trouverait $g(x, y)$ moyennant des relations de la forme (1.6). On observe que $g_{xx}(x, y)$ est une fonction spline du troisième degré par rapport à y. Résolvons $m + 1$ problèmes à une variable pour $y = y_j$ et $g_{xx}(x, y)$ dont on connaît les valeurs aux nœuds de D_h. On trouve finalement $g_{xxyy}(x, y)$ définie sur D_h. Avec la notation $K_{ij} = g_{xxyy}(x_i, y_j)$, on a

$$g_{xx}(x_{i-1}, y) = K_{i-1\,j-1}\frac{(y_j - y)^3}{6\tau_j} + K_{i-1\,j}\frac{(y - y_{j-1})^3}{6\tau_j} +$$
$$+ \left(M_{i-1\,j-1} - \frac{K_{i-1\,j-1}\tau_j^2}{6}\right)\frac{y_i - y}{\tau_j} + \left(M_{i-1\,j} - \frac{K_{i-1\,j}\tau_j^2}{6}\right)\frac{y - y_{j-1}}{\tau_j}, \quad (2.7)$$

$$g_{xx}(x_i, y) = K_{i\,j-1}\frac{(y_j - y)^3}{6\tau_j} + K_{ij}\frac{(y - y_{j-1})^3}{6\tau_j} +$$
$$+ \left(M_{ij} - \frac{K_{i\,j-1}\tau_j^2}{6}\right)\frac{y_j - y}{\tau_j} + \left(M_{ij} - \frac{K_{ij}\tau_j^2}{6}\right)\frac{y - y_{j-1}}{\tau_j}. \quad (2.8)$$

On obtient facilement $g(x, y)$ par recours à (1.6) :

$$g(x, y) = g_{xx}(x_{i-1}, y)\frac{(x_i - x)^3}{6h_i} + g_{xx}(x_i, y)\frac{(x - x_{i-1})^3}{6h_i} +$$
$$+ \left(g(x_{i-1}, y) - \frac{g_{xx}(x_i, y)\,h_i^2}{6}\right)\frac{x_i - x}{h_i} +$$
$$+ \left(g(x_i, y) - \frac{g_{xx}(x_i, y)\,h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}. \quad (2.9)$$

Quel volume de calcul effectue-t-on pour déterminer la fonction spline bicubique $g(x, y)$? Avant de calculer ses valeurs aux points

concernés, on résout $(n + 1) + (m + 1) + (m + 1) = 2m + n + 3$ systèmes algébriques linéaires du type (1.9) et on cherche N_{ij}, M_{ij}, K_{ij}, $i = 0, \ldots, n$; $j = 0, \ldots, m$. La recherche de $g(x, y)$ en un point du domaine exige qu'on calcule 5 fois une formule (1.6) (notamment (2.5), (2.6), (2.7), (2.8), (2.9)). On pourrait certes utiliser cette chaîne de formules pour obtenir $g(x, y)$ sous forme polynomiale explicite dans chaque cellule. Mais les tableaux des coefficients du polynôme occuperont environ 16 nm éléments de la mémoire interne contre $4nm$ dans l'algorithme considéré encore que le calcul de la fonction en un point demande environ quatre fois plus d'opérations.

On peut intervertir les variables et aborder en conséquence $2n + m + 3$ problèmes unidimensionnels sans que le résultat change pas plus que le nombre total d'opérations.

L'algorithme décrit se généralise sans peine à des domaines de plusieurs dimensions du type parallélépipède. Pour les résultats correspondants, consulter Zavialov [XIV; 5].

3.3. Approximation r-régulière de fonctions de plusieurs variables

Ce numéro constitue une approche, parmi tant d'autres, du problème de prolonger les fonctions de plusieurs variables définies avec une erreur sur des réseaux chaotiques. On s'intéressera en premier lieu à l'aspect algorithmique de la méthode proposée car la théorie correspondante est pour le moment heuristique.

La méthode est due à Tsétsokho, Bélonossov, Bélonossova [XIV; 11] (pour une modification à peine différente voir Lébédev [XIV; 6], et part de la notion de partition de l'unité très connue en analyse mathématique.

On dit qu'une suite (φ_i) de fonctions de classe $C^r(V)$, $V \subset R^m$ étant un ouvert, constitue une partition de l'unité sur $\Omega \subset V$ si $0 \leqslant \varphi_i(x) \leqslant 1$ et $\sum_i \varphi_i(x) = 1$ quel que soit $x \in \Omega$.

On assujettit d'ordinaire φ_i à s'annuler en dehors d'un ouvert $V_i \subset V$ donné. L'ensemble (V_i) constitue évidemment un recouvrement de Ω et la partition de l'unité (φ_i) est dite subordonnée à ce recouvrement.

Dans les applications, on utilise naturellement des recouvrements localement finis par des ensembles standards, à savoir par des boules et intervalles m-dimensionnels. On en choisit les centres et les diamètres selon les propriétés de la fonction à approcher f: $\Omega \to R$. Dans la suite, nous considérerons les recouvrements par des cubes de dimension m. On procède comme suit pour former les partitions de l'unité subordonnées.

Notons $\xi^{(i)}$ le centre d'un cube V_i et $\delta^{(i)}$ la longueur de l'arête. On se donne une fonction (standard) $\psi\colon R^m \to R$ de classe $C^r(R^m)$ telle que

$$\psi(x) > 0 \quad \text{si} \quad x \in \left(-\frac{1}{2}, \frac{1}{2}\right) \times \ldots \times \left(-\frac{1}{2}, \frac{1}{2}\right) = V_0,$$

$$\psi(x) = 0 \quad \text{si} \quad x \notin V_0.$$

On pose par exemple (pour $x = (x_1, \ldots, x_m) \in V_0$)

$$\psi(x) = \prod_{i=1}^{m} \cos^{r+1}(\pi x_i) \text{ ou } \psi(x) = \prod_{i=1}^{m} \left(\frac{1}{4} - x_1^2\right)^{r+1},$$

et ainsi de suite. Posons

$$\psi_i(x) = \begin{cases} \psi\left(\dfrac{x - \xi^{(i)}}{\delta^{(i)}}\right), & x \in V_i, \\ 0, & x \notin V_i, \end{cases}$$

quel que soit i. Les fonctions

$$\varphi_i(x) = \frac{\psi_i(x)}{\sum\limits_j \psi_j(x)}$$

sont de classe $C^r(V)$, $V = \bigcup\limits_i V_i$, et forment sur Ω une partition de l'unité subordonnée au recouvrement (V_i).

Admettons qu'on associe à $f\colon \Omega \to R$ une suite $(f_i)\colon V \to R$ de fonctions de classe $C^r(V)$ telle que chaque f_i soit « proche » de f sur V_i. Soit

$$\| (f - f_i) \mid_{V_i} \| \leqslant \varepsilon$$

pour la norme uniforme quel que soit i. Alors $\tilde{f}$ définie par

$$\tilde{f}(x) = \sum_i f_i(x)\, \varphi_i(x), \qquad x \in \Omega, \tag{3.1}$$

est « proche » de f sur Ω. En effet, l'identité

$$f(x) = \sum_i f_i(x)\, \varphi_i(x), \qquad x \in \Omega,$$

entraîne pour tout $x \in \Omega$

$$|f(x) - \tilde{f}(x)| = \left|\sum_i (f(x) - f_i(x))\, \varphi_i(x)\right| \leqslant$$

$$\leqslant \max_i \sup_{x \in V_i} |f(x) - f_i(x)| \sum_i \varphi_i(x) \leqslant \varepsilon,$$

i.e.

$$\|f - \tilde{f}\| \leqslant \varepsilon.$$

Prolongeons la fonction f moyennant une partition de l'unité; si elle est définie aux nœuds d'un réseau ($x^{(j)} \in \Omega$), on l'approche localement pour chaque i par $f_i(x)$ basée sur un ensemble de nœuds ($x^{(j)}$) voisins du centre $\xi^{(i)}$ du cube V_i, puis on « recolle » ces approximations par la formule (3.1). Dans la suite, on suppose que les nœuds $x^{(i)}$ coïncident avec les centres $\xi^{(i)}$ des cubes V_i et que les approximations locales sont des fonctions constantes égales à $f(x^{(i)})$.

Evaluons l'erreur de discrétisation dans le cas particulier où

$$f(x) = x, \qquad x \in R;$$

le réseau est régulier infini à nœuds $x_k = kh$, $k = 0, \pm 1, \pm 2, \ldots$; les centres $\xi^{(k)}$ des intervalles V_k coïncident avec les nœuds et $\delta^{(k)} = (2N+1)h$, N étant un entier naturel. Admettons que la partition de l'unité (φ_k) est de classe $C^r(R)$, $r \geqslant 3$, et que $\varphi_0(x)$ est paire.

Dans ce cas le reste

$$\sigma(x) = x - \sum_k x_k \varphi_k(x)$$

constitue une fonction impaire périodique, de période h, qui vaut 0 aux nœuds entiers et demi-entiers. On se borne donc aux cas: 1) $0 \leqslant x \leqslant \frac{h}{4}$; 2) $\frac{h}{4} \leqslant x \leqslant \frac{h}{2}$.

1) Comme $\varphi_k(x) = \varphi_0(x - x_k)$ et φ_0 est paire, on a

$$\widetilde{x} = \sum_k x_k \varphi_k(x) = \sum_k x_k \varphi_0(x - x_k) = \sum_k x_k [\varphi_0(x_k - x) - \varphi_0(x_k)].$$

D'où par la formule de Taylor

$$\widetilde{x} = -x \sum_k x_k \varphi_0'(x_k) + \frac{x^2}{2} \sum_k x_k \varphi_0''(x_k) - \frac{x^3}{6} \sum_k x_k \varphi_0'''(x_k - \theta_k x).$$

Le deuxième terme à droite est évidemment nul, si bien que

$$\widetilde{x} = -x \sum_k x_k \varphi_0'(x_k) - \frac{x^3}{6} \sum_k x_k \varphi_0'''(x_k - \theta_k x). \tag{3.2}$$

Transformons

$$\bar{x} = -x \sum_k x_k \varphi_0'(x_k)$$

par Abel, il vient

$$\bar{x} = xh \sum_{k=-N}^{N-1} \sum_{r=-N}^{k} \varphi_0'(rh) + xx_N \sum_{r=-N}^{k} \varphi_0'(rh) = xh \sum_{k=-N}^{N-1} \sum_{r=-N}^{k} \varphi_0'(rh).$$

Mais

$$h \sum_{r=-N}^{k} \varphi_0'(rh) = \int_{-(N+1/2)h}^{(k+1/2)h} \varphi_0'(x)\,dx + \sigma_1(x) = \varphi_0\left(x_k + \frac{h}{2}\right) + \sigma_1(k),$$

où

$$|\sigma_1(k)| \leqslant (\max_x |\varphi_0'''(x)|)\frac{h^3}{24}(k+N+1). \tag{3.3}$$

Ainsi,

$$\bar{x} = x \sum_k \varphi_0\left(x_k + \frac{h}{2}\right) + x \sum_{k=-N}^{N-1} \sigma_1(x) = x + x \sum_{k=-N}^{N-1} \sigma_1(k). \tag{3.4}$$

Les formules (3.2) à (3.4) donnent l'estimation

$$|\sigma(x)| \leqslant (\max_x |\varphi_0'''(x)|)\left[\frac{x^3 h}{6} \sum_{k=-N}^{N} k + \frac{xh^3}{24} \sum_{k=-N}^{N-1} (k+N+1)\right] \leqslant$$

$$\leqslant (\max_x |\varphi_0'''(x)|)\, h^4 \frac{3}{2^7} N\left(N + \frac{5}{9}\right) =$$

$$= C \frac{h^4 3\left(N+\frac{5}{9}\right)N}{2^7 (2N+1)^3 h^3} \leqslant C \frac{h}{288N}. \tag{3.5}$$

La constante C est la valeur de la dérivée troisième de plus grand module de $x(y) = \varphi_0(y(2N+1)h)$, elle décroît avec la croissance de N.

On obtient le même résultat dans le cas 2).

L'estimation ci-dessus constitue un critère de choix de la fonction initiale ψ qui détermine la partition de l'unité, à savoir : la dérivée troisième de ψ doit être le plus petite possible.

On aboutit à des estimations analogues pour une fonction linéaire de plusieurs variables.

On se pose maintenant le problème fondamental de ce numéro.

Etant donnée une fonction $f\colon \Omega \to R$ connue à une quantité ε près (en norme uniforme) sur un réseau irrégulier $(x^{(i)})$, $i = 1, 2, \ldots, n$, on demande une fonction $\tilde{f}$ de classe C^r telle qu'on ait la condition d'approximation

$$\|f - \tilde{f}\| = \max_{i=1,\ldots,n} |f(x^{(i)}) - \tilde{f}(x^{(i)})| \leqslant \varepsilon \tag{3.6}$$

et la condition de lissage heuristique qui exige que les dérivées d'un certain ordre de $\tilde{f}$ soient « non oscillantes ».

La méthode ci-dessus procède du prolongement répétitif décrit qui utilise une suite de partitions de l'unité. L'algorithme se déroule

par des relations qui garantissent à notre avis la condition de lissage (au moins au sens heuristique indiqué).

Désignons par L_0 l'opérateur (d'approximation lissante) de l'espace des fonctions discrètes $\{f\colon x^{(i)} \to R\}$ dans l'espace $C^r(V)$ des fonctions dérivables:

$$L_0(f)(x) = \sum_{i=1}^{n} f(x^{(i)})\,\varphi_i(x), \tag{3.7}$$

où (φ_i) est une partition de l'unité sur Ω du type décrit, les centres des supports des φ_i étant les nœuds $x^{(i)}$ du réseau.

Evaluons la norme de l'écart:

$$\|f - L_0(f)\| = \max_{j=1,\ldots,n} \Big| f(x^{(j)}) - \sum_{i=1}^{n} f(x^{(i)})\,\varphi_i(x^{(j)}) \Big| =$$

$$= \min_{j=1,\ldots,n} \Big| f(x^{(j)})\,(1-\varphi_j(x^{(j)})) - \sum_{\substack{i=1\\ i\neq j}}^{n} f(x^{(i)})\,\varphi_i(x^{(j)}) \Big| \leqslant$$

$$\leqslant 2\,\|f\|\,[\max_{j=1,\ldots,n} \varphi_j(x^{(j)})]. \tag{3.8}$$

Il en découle qu'avec (3.7) appliquée une fois, on vérifie la condition d'approximation si l'on choisit une partition de l'unité telle que

$$\lambda = \min_{j=1,\ldots,n} \varphi_j(x^{(j)}) \tag{3.9}$$

soit suffisamment proche de 1. Mais on a ce dernier résultat si ou bien les ensembles V_i contiennent peu de nœuds, ou bien les fonctions φ_j sont « trop en delta », ce qui compromet, selon l'estimation (3.5), les propriétés de lissage de L_0.

On voit qu'une seule application de l'opérateur L_0 ne conduit en général pas à la solution. On procède donc comme suit.

Soit $\{\varphi_i^{(k)}\}_{k=1,2,\ldots}$ une suite de partitions de l'unité qu'on suppose, pour fixer les idées, subordonnées à un même recouvrement de Ω et soit $(L_0^{(k)})$ la suite associée d'opérateurs d'approximation lissante qu'on définit pour chaque $(\varphi_i^{(k)})$ par la formule (3.7). On cherche $\tilde{f}$ sous la forme

$$\tilde{f} = \sum_{k=1}^{k_0} L_0^{(k)}(f^{(k-1)}), \tag{3.10}$$

avec un k_0, $(f^{(k)})$ étant une suite de fonctions discrètes définies par

$$f^{(k)}(x^{(j)}) = \begin{cases} f(x^{(j)}), & k=0, \\ f^{(k-1)}(x^{(j)}) - L_0^{(k)}(f^{(k-1)})(x^{(j)}), & k>0, \end{cases} \tag{3.11}$$

$j = 1, 2, \ldots, n$. Sommons sur $k = 0, 1, \ldots, k_0$, il vient

$$f = \tilde{f} + f^{(k_0)},$$

et pour que la condition d'approximation ait lieu, il faut que $\|f^{(k_0)}\|$ ne dépasse pas ε. Cette condition nécessaire est toujours vérifiée pour un choix spécial de la suite $\{\varphi_i^{(k)}\}$. Supposons, en effet, qu'on a, à partir d'un k_1,

$$\lambda^{(k)} = \min_{j=1,\ldots,n} \varphi_j^{(k)}(x^{(j)}) \geqslant \sigma > \frac{1}{2}$$

quel que soit $k \geqslant k_1$. Dans ce cas,

$$\|E - L_0^{(k)}\| \leqslant 2(1-\sigma) < 1$$

par suite de (3.8), donc $\|f^{(k)}\| \xrightarrow[k\to\infty]{} 0$.

Avec une suite $\{\varphi_i^{(k)}\}$ telle que les quantités $\lambda^{(k)}$ soient petites au possible si $(\|f^{(k)}\|)$ décroît de façon monotone à une vitesse $v > 1$ fixée à l'avance (i.e. $\|f^{(k)}\|/\|f^{(k+1)}\| \leqslant v$), on arrive également à satisfaire la condition de lissage (parce que la partie « non lisse » de la somme (3.10) est due aux valeurs $f^{(k)}$ petites en norme).

Dans la pratique, il y a intérêt à procéder comme suit. Si $\psi^{(k)}(x)$ est une fonction standard pour la partition $(\varphi_i^{(k)})$, son homologue pour la partition $(\varphi_i^{(k+1)})$ est choisi sous la forme

$$\psi^{(k+1)}(x) = \psi^{(k)}(x)\, g^{(k)}(x),$$

où le facteur $g^{(k)}(x)$ est ou bien

$$g^{(k)}(x) = \exp[-\mu_k (x_1^2 + x_2^2 + \ldots + x_m^2)], \qquad \mu_k \geqslant 0,$$

ou bien

$$g^{(k)}(x) = [1 + \mu_k (x_1^2 + x_2^2 + \ldots + x_m^2)]^{-1}, \qquad \mu_k \geqslant 0,$$

et ainsi de suite.

On organise le calcul de façon que $\mu_k = 0$ si $\|f^{(k-1)}\|/\|f^{(k)}\| \geqslant v$ et $\mu_k = \mu > 0$ dans le cas contraire.

Ainsi, l'algorithme décrit renferme trois paramètres régulateurs :

1) quantité $\lambda^{(0)} = \lambda$ définie par (3.9) ;

2) vitesse $x > 1$ de diminution du résidu $\|f^{(k)}\|$;

3) μ qui règle la vitesse à laquelle s'accentue la « forme en delta » des partitions de l'unité.

On conçoit qu'en optimisant par rapport à ces paramètres on arrive à minimiser les fonctionnelles dont les valeurs déterminent diverses notions de dérivées « oscillantes » de $\tilde{f}$.

REMARQUE. Si la suite de partitions de l'unité a ses termes (φ_i) identiques, on trouve, en vertu de (3.10), (3.11),

$$\tilde{f}(x) = \sum_{i=1}^{n} \Big(\sum_{k=1}^{k_0} f^{(k-1)}(x^{(i)})\, \varphi_i(x) \Big).$$

Une question se pose : est-ce qu'on peut utiliser un seul ensemble de (φ_i) et chercher la solution $\tilde{f}$ sous la forme

$$\tilde{f}(x) = \sum_{i=1}^{n} c_i \varphi_i(x), \tag{3.12}$$

avec c_i définis par

$$\left\| f - \sum_{i=1}^{n} c_i \varphi_i(x) \right\| \leqslant \varepsilon \tag{3.13}$$

sous l'hypothèse d'être dans la condition de lissage ?

Cette façon de raisonner est fausse par essence car $f(x^{(i)})$ sont données avec une erreur.

En effet, si

$$\det \| \varphi_i(x^{(j)}) \| \neq 0, \tag{3.14}$$

alors on vérifie (3.13) pour $\varepsilon = 0$, auquel cas les erreurs sur $f(x^{(i)})$ « mesurées » se propagent à (3.12). Aussi il y a intérêt à réaliser au possible l'estimation (3.13), ce qui exige naturellement la condition

$$\max_{i=1, \ldots, n} | f(x^{(i)}) - c_i | \leqslant E, \quad E > 0, \tag{3.15}$$

qui limite la norme $\| f - c \|$. On ne possède malheureusement aucun critère de choix de E, et si cette quantité est donnée quelconque (il est naturel qu'on se la donne « petite »), on est conduit à un problème qui est en général impossible (sans parler du fait que (3.14) n'est pas nécessairement remplie).

L'algorithme décrit provient justement du caractère contradictoire des conditions d'approximation et de lissage et de (3.15).

3.4. Eléments de la théorie générale des fonctions splines

Les fonctions splines du troisième degré interviennent, on l'a montré, dans le problème de minimisation de la fonctionnelle quadratique $\int_a^b [u''(x)]^2 \, dx$, analogue de l'énergie de flexion d'une tige élastique. Ce fait a déterminé deux grandes tendances de la théorie des fonctions splines.

La première tendance est de remplacer l'opérateur de dérivation d'ordre deux par un opérateur différentiel général L à coefficients variables, voire discontinus, si bien que la fonctionnelle à minimiser s'écrit $\int_a^b [Lu(x)]^2 \, dx$. Cet artifice rend la théorie des fonctions splines très maniable et permet de l'utiliser pour approcher des fonctions avec minimisation de l'énergie du phénomène physique concret.

La seconde tendance consiste à renoncer à cette situation simple qu'est une courbe fixée aux points du réseau, au profit d'autres fonctionnelles linéaires qui semblent plus naturelles dans tel ou tel cas. Il s'agit des splines de degré pair, des fonctions splines trigonométriques et d'autres êtres mathématiques qui sont très employés aujourd'hui. On assiste en même temps au développement de la théorie des splines à plusieurs variables. Cette approximation de fonctions est doublée de lissage.

C'est à Atteia [XIV] qu'on doit la première définition générale des fonctions splines d'interpolation et de lissage (d'ajustement). Soient X, Y deux espaces de Hilbert et T: $X \to Y$ un opérateur linéaire borné de X dans Y. On définit sur X un système de fonctionnelles linéaires bornées k_i, $i = 1, \ldots, n$, qu'on suppose être linéairement indépendantes. On appelle *fonction spline d'interpolation* un élément $\sigma \in X$ vérifiant deux conditions

$$\begin{aligned} &1) \qquad (k_i, \sigma)_X = r_i, \quad i = \overline{1, n}, \\ &2) \qquad (T\sigma, T\sigma)_Y = \| T\sigma \|_Y^2 = \min. \end{aligned} \tag{4.1}$$

Les symboles $(,)_X$, $(,)_Y$ désignent les produits scalaires dans X et Y; r_i, $i = \overline{1, n}$, sont des nombres donnés à l'avance. On appelle *fonction spline de lissage* (*d'ajustement*) un élément $\sigma_\alpha \in X$ qui réalise le minimum de la fonctionnelle

$$\Phi_\alpha(u) = \alpha \| Tu \|_Y^2 + \sum_{i=1}^{n} [(k_i, u) - r_i]^2, \quad \alpha > 0. \tag{4.2}$$

On doit à Atteia des théorèmes généraux d'existence et d'unicité pour les problèmes (4.1) et (4.2), qui permettent de construire toutes sortes d'approximations par des fonctions splines. L'ouvrage fondamental d'Anselone et Laurent [XIV] expose un algorithme général de formation de fonctions splines des deux types. On montre que les problèmes (4.1) et (4.2) se ramènent en fin de compte à la résolution de certains systèmes algébriques à matrices symétriques définies positives.

La convergence des fonctions splines a intéressé de nombreux auteurs qui l'ont généralisée à divers degrés (Ahlberg, Nilson, Walsh [III], Schoenberg [XIV], Varga [I], Morozov [XIV; 8], Zavialov [XIV; 5], Vassilenko [XIV; 2]). On trouve dans Varga [I] une bonne analyse des estimations de la convergence sur différentes classes fonctionnelles.

Nous donnerons plus loin plusieurs théorèmes de convergence assez généraux qui s'appliquent aussi bien au cas classique de fonctions splines basées sur les réseaux rectangulaires qu'aux cas plus complexes tels que les réseaux chaotiques de plus en plus fins et les

fonctions splines construites moyennant un système arbitraire de fonctionnelles. Ces théorèmes sont l'œuvre de Joly [XIV], Morozov [XIV; 8] et Vassilenko [XIV; 2].

Commençons par la convergence des fonctions splines d'interpolation. Soient X un espace de Hilbert et φ^* son élément à approximer. Considérons dans X un système de vecteurs de base $k_1, k_2, \ldots, \ldots, k_n, \ldots$ qui n'est en général pas orthogonal. Si $T: X \to Y$ est un opérateur linéaire borné de X dans un autre espace hilbertien Y, alors on cherche l'approximant sous forme de fonction spline σ_n :

$$\begin{aligned} &(\sigma_n, k_i)_X = (\varphi^*, k_i)_X = r_i, \qquad i = 1, \ldots, n, \\ &\| T\sigma_n \|_Y^2 = \min. \end{aligned} \tag{4.3}$$

Sous l'hypothèse qu'il y a existence et unicité de σ_n pour un $n = n_0$, les fonctions splines σ_n existent quel que soit $n \geqslant n_0$. On se demande si σ_n convergent vers le vecteur exact φ^*. Selon Vassilenko [XIV; 2], cette convergence a lieu et, chose beaucoup plus importante dans la pratique, la suite $T\sigma_n$ tend vers $T\varphi^*$ pour la norme de Y si l'on est dans les conditions du théorème d'existence et d'unicité et si le noyau de l'opérateur T est de dimension finie.

Il y a intérêt à utiliser ces théorèmes généraux afin d'évaluer la convergence en fonction de la diminution du pas du réseau. S'agissant des splines formées de morceaux de polynômes du degré $2n - 1$ qui se raccordent sur le segment $[a, b]$, l'approximation de $\varphi^* \in W_2^n$ et de ses dérivées par σ_N s'évalue par

$$\begin{aligned} &\| \sigma_N^{(i)} - \varphi^{*(i)} \|_{C[a, b]} = o\left(h^{n-i-1/2}\right), \qquad i = 0, 1, \ldots, n-1, \\ &\| \sigma_N^{(n)} - \varphi^{*(n)} \|_{L_2[a, b]} = o(1). \end{aligned} \tag{4.4}$$

La théorie générale de la convergence peut être utile lorsque toutes les méthodes connues tombent en défaut. C'est le cas des fonctions splines basées sur les points d'interpolation répartis de façon chaotique dans un domaine bidimensionnel. On forme les fonctions splines dans des domaines de dimension 2 à partir de la condition de minimisation de la fonctionnelle

$$\Phi(u) = \int_\Omega \left[\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 u}{\partial x \, \partial y} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 \right] d\Omega \tag{4.5}$$

si u prend des valeurs données dans un ensemble ponctuel fixe. Si $\varphi^* \in W_2^2(\Omega)$, on obtient par un théorème général de convergence

$$\| \sigma_N - \varphi^* \|_{C[\Omega_\delta]} = o\left(h^{3/2}\right), \tag{4.6}$$

où σ_N est une fonction spline ; h le paramètre du resserrement du réseau chaotique, qui est défini comme le paramètre d'un ε-réseau formé de points d'interpolation de Ω ; Ω_δ tout δ-intérieur de Ω, i.e. l'ensemble des points $\in \Omega$ dont la distance à la frontière dépasse δ.

Voici, dans ses traits généraux, la théorie de la convergence des fonctions splines d'interpolation. Mais la recherche d'une fonction spline d'interpolation se ramène à la minimisation d'une fonctionnelle quadratique soumise à des contraintes linéaires, tandis que le cas d'une fonction spline de lissage constitue un problème de minimisation sans contraintes qui s'avère parfois plus facile. Morozov [XIV; 8] a montré que la fonction spline de lissage σ_α converge, lorsque $\alpha \to 0$, vers la fonction spline d'interpolation σ pour la norme de X.

La génération de fonctions splines d'interpolation et d'ajustement par l'algorithme théorique universel d'Anselone et Laurent [XIV] s'avère quasi impossible dans la pratique dès la dimension 2, et on s'adresse donc à des méthodes d'approximation. On procède comme suit. On substitue au problème de rechercher σ le problème de former σ_α définie en minimisation sur X tout entier de la fonctionnelle

$$\Phi(u) = \alpha \| Tu \|_X^2 + \sum_{i=1}^{n} [(k_i, u) - r_i]^2, \qquad \alpha > 0, \tag{4.7}$$

avec α petit. On considère dans X un sous-espace E_h de dimension finie dont les éléments approchent ceux de X avec la précision désirée, et on minimise $\Phi_\alpha(u)$ sur E_h. Le point de minimum constitue une *fonction spline relative à* E_h. Pour la convergence de ces fonctions splines voir Vassilenko [XIV; 2].

Cette tactique s'avère très fructueuse dans le problème suivant: soit Ω un domaine rectangulaire à points d'interpolation P_i, $i = 1, \ldots, s$, irrégulièrement espacés. On demande une fonction d'interpolation régulière dans Ω. Superposons à Ω un réseau rectangulaire Ω_h et associons-le à l'espace E_h de dimension finie des interpolants réguliers. Nous savons construire ces espaces dont les éléments peuvent être des *fonctions splines bicubiques* ou des *prolongements de Riabenki* de classe de régularité C^2. On cherche sur E_h le minimum de la fonctionnelle

$$\Phi_\alpha(u) = \alpha \int_\Omega [u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2]\, d\Omega + \sum_{i=1}^{s} [u(P_i) - r_i], \tag{4.8}$$

avec $\alpha > 0$ petit. Ce problème conduit à un système algébrique linéaire. Si l'on choisit dans E_h une base de fonctions à support borné, le système est à matrice bande. On prend pour fonctions de base des B-splines du troisième degré ou, ce qui est mieux encore à notre avis, des fonctions de base locales intervenant dans l'interpolation au sens de Riabenki. Pivovarova et Poukhnatchéva [XIV; 9] ont élaboré un programme intéressant qui réalise un processus avec choix simultané du paramètre α.

Il nous reste à nous prononcer sur le choix du paramètre de lissage dont l'importance pratique est incontestable. Parmi des critères qui le régit, le « critère des résidus normés » paraît naturel pour un large éventail de problèmes (voir tout particulièrement Morozov [XIV; 8]).

Plaçons-nous dans un problème de lissage, i.e. on cherche un élément σ_α de X qui minimise la fonctionnelle

$$\Phi_\alpha(u) = \alpha \| Tu \|_Y^2 + \sum_{i=1} [(k_i, u) - r_i]^2 \equiv \alpha \| Tu \|_Y^2 + \| Ku - r \|^2. \quad (4.9)$$

L'opérateur $K: X \to R^n$ agit par la formule

$$Ku = \{(u, k_1), \ldots, (u, k_n)\}, \quad (4.10)$$

et $r = (r_1, r_2, \ldots, r_n)$ est un vecteur donné. Le critère mentionné consiste à choisir α moyennant la condition

$$\| K\sigma_\alpha - r \| = \varepsilon, \quad (4.11)$$

ε fixe étant le « niveau acceptable du résidu ».

Introduisons les notations

$$\varphi(\alpha) = \| K\sigma_\alpha - r \|, \qquad 0 < \alpha < +\infty. \quad (4.12)$$

On résout l'équation non linéaire en α:

$$\varphi(\alpha) = \varepsilon.$$

Faisons la substitution $\alpha = 1/\beta$ et considérons l'équation

$$\varphi(\alpha) = \varphi(1/\beta) = \psi(\beta) = \varepsilon \quad (4.13)$$

par rapport à β. La fonction $\psi(\beta)$ est strictement monotone décroissante et $\psi(\beta) = \varepsilon$ admet une solution unique pour $\varepsilon < \varepsilon_0$, et $\varepsilon_0 = \| Ke - r \|$ avec e la solution du problème des moindres carrés

$$\| Ke - r \| = \min_{u \in \mathrm{Ker}\, T} \| Ku - r \|$$

sur le noyau de T. Il est naturel de résoudre (4.13) par la méthode de Newton. Gordonova et Morozov [XIV; 4] proposent d'élever les deux membres à une puissance réelle q,

$$\psi^q(\beta) = \varepsilon^q, \quad (4.14)$$

et d'accélérer au possible à l'aide de q la convergence du processus itératif de Newton. Pour qu'il y ait convergence pour toute approximation initiale, q doit être compris dans l'intervalle $-1 \leqslant q < 0$, $q > 0$, la plus grande vitesse étant garantie par $q = -1$. Ainsi, il faut résoudre l'équation

$$f(\beta) = 1/\varphi(1/\beta) = \varepsilon^{-1}. \quad (4.15)$$

On effectue les itérations de Newton par la formule

$$\beta^{p+1} = \beta^p - \frac{f(\beta^p)}{f'(\beta^p)} = \beta^p - (\beta^p)^2 \frac{\varphi(1/\beta^p)}{\varphi'(1/\beta^p)}. \tag{4.16}$$

Quel est le volume de calcul par pas? Le calcul de $\varphi(1/\beta^p) = \varphi(\alpha^q)$ demande la résolution d'un problème de lissage par des fonctions splines pour $\alpha = \alpha^p$ fixe, i.e. on a à résoudre un système algébrique linéaire. On doit de plus connaître $\varphi'(\alpha^p)$ et on résout à cette fin le même système algébrique linéaire avec un autre second membre (voir Gordonova, Morozov [XIV; 4]).

La théorie générale du lissage est visiblement équivalente à une méthode de régularisation de problèmes mal posés. On admet par exemple que K est une matrice dégénérée et la fonction spline d'interpolation u constitue la solution normale du système compatible

$$\begin{aligned} Ku &= f, \\ \| u \| &= \min. \end{aligned} \tag{4.17}$$

La solution (fonction spline d'interpolation) est approchée pour $\alpha \to 0$ par la solution u_α (fonction spline de lissage) du système régularisé

$$\begin{aligned} &\alpha \| u_\alpha \|^2 + \| Ku_\alpha - f \|^2 = \min_u, \\ &(\alpha I + K^*K)\, u_\alpha = K^*f. \end{aligned} \tag{4.18}$$

CHAPITRE 4

MÉTHODES DE RÉSOLUTION DES PROBLÈMES STATIONNAIRES

La résolution des problèmes stationnaires constitue un chapitre plus ou moins autonome de l'Analyse numérique encore que dans de nombreux cas à opérateurs positifs la solution s'interprète comme solution limite pour $t \to \infty$ d'un problème non stationnaire. Dans la stationnarisation asymptotique des problèmes stationnaires, on ignore délibérément les valeurs intermédiaires de la solution car elles sont sans importance, tandis que ces valeurs ont une signification physique pour les problèmes non stationnaires. Cette circonstance unit, tout en les différenciant, ces classes de problèmes. Illustrons-le sur un exemple.

Soit le problème

$$A\varphi = f,$$

où

$$A > 0, \ \varphi \in \Phi \text{ et } f \in F.$$

Considérons le problème non stationnaire *)

$$\frac{\partial \psi}{\partial t} + A\psi = f,$$

$$\psi = 0 \quad \text{pour} \quad t = 0$$

et définissons φ, ψ et f par les formules

$$\varphi = \sum_n \varphi_n u_n, \qquad \psi = \sum_n \varphi_n u_n, \qquad f = \sum_n f_n u_n,$$

où

$$Au_n = \lambda_n u_n, \qquad A^* u_n^* = \lambda_n u_n^*,$$

$\varphi_n = (\varphi, u_n^*)$, $\psi_n = (\psi, u_n^*)$, $f_n = (f, u_n^*)$ et $\{u_n\}$ et $\{u_n^*\}$ sont des bases biorthogonales. Il est classique de se ramener aux problèmes respectifs

$$\lambda_n \varphi_n = f_n$$

et

$$\frac{d\psi_n}{dt} + \lambda_n \psi_n = f_n, \qquad \psi_n(0) = 0,$$

*) Nous omettrons de spécifier les domaines des opérateurs.

et d'obtenir les coefficients de Fourier

$$\varphi = \sum_n \frac{f_n}{\lambda_n} u_n, \qquad \psi = \sum_n \frac{f_n}{\lambda_n} (1 - e^{-\lambda_n t}) u_n.$$

Le spectre de A étant supposé réel, on a $\lambda_n > 0$, $n = 1, 2, \ldots$, d'où

$$\lim_{t \to \infty} \psi = \varphi.$$

Si l'opérateur du problème stationnaire possède un spectre de structure quelconque, il se peut que cette relation simple et nette entre les solutions fasse défaut.

Il va sans dire que le problème non stationnaire pour ψ obtenue se prête à des méthodes aux différences par rapport à t. Par exemple

$$\frac{\psi^{j+1} - \psi^j}{\tau} + A\psi^j = f.$$

On a

$$\psi^{j+1} = \psi^j - \tau (A\psi^j - f).$$

Si c'est le problème stationnaire qu'on aborde, alors une relation déterminée entre τ et β (A) entraîne

$$\lim_{j \to \infty} \psi^j = \varphi.$$

Le paramètre τ est fonction ou non de j. Dans les deux cas, on ferait bien de considérer j comme étant un indice repérant non pas un pas de temps, mais une itération.

Voici une autre différence entre les deux classes de problèmes: dans les problèmes non stationnaires, on garantit l'exactitude de la solution avec τ suffisamment petits, tandis que dans les cas stationnaires les paramètres τ optimaux sont choisis à partir de la condition de minimum du nombre d'itérations et ils peuvent être assez élevés.

Pour les méthodes itératives en algèbre linéaire voir les monographies de Faddeev et Faddeeva (VIII; 16, 22], Forsythe, Wasow [III], Householder [III], Warga [III], Marchouk, Lébédev (XVII; 12], Samarski [III; 12], Bakhvalov [III], Marchouk, Kouznétsov [VIII; 12].

4.1. Eléments de la théorie des méthodes itératives

Dans tout le chapitre, l'opérateur A sera une matrice carrée, si bien qu'on aura, au lieu du problème initial, un système algébrique linéaire. Ceci étant, les vecteurs et les matrices seront toujours à valeurs réelles et la matrice A sera non dégénérée (sauf au numéro 4.5). Soit donc le système

$$A\varphi = f, \tag{1.1}$$

avec A une matrice et φ et f deux vecteurs.

La plupart des méthodes itératives pour les systèmes linéaires sont gouvernées par la formule

$$B_j \frac{\varphi^{j+1} - \varphi^j}{\tau_j} = -(A\varphi^j - f), \tag{1.2}$$

$\{B_j\}$ étant une suite de matrices non dégénérées et $\{\tau_j\}$ une suite de paramètres réels. Avec la notation $H_j = \tau_j B_j^{-1}$, le processus (1.2) s'écrit de façon équivalente

$$\varphi^{j+1} = \varphi^j - H_j(A\varphi^j - f). \tag{1.3}$$

Les vecteurs $\xi^j = A\varphi^j - f$ constituent les *vecteurs résidus* de la méthode itérative (1.2) et $\psi^j = \varphi^j - \varphi^*$ ($\varphi^* = A^{-1}f$ est la solution exacte du système (1.1)) en sont les *vecteurs erreurs*. En retranchant de deux membres de (1.3) le vecteur φ^* et en faisant la substitution $f = A\varphi^*$, on obtient pour la suite de vecteurs erreurs

$$\psi^{j+1} = T_j \psi^j, \tag{1.4}$$

la matrice

$$T_j = E - H_j A \tag{1.5}$$

étant l'*opérateur de transition* de la méthode considérée. Multiplions (1.4) par la matrice A, il vient la relation pour la suite de vecteurs résidus

$$\xi^{j+1} = (E - AH_j)\,\xi^j. \tag{1.6}$$

La méthode itérative (1.2) est dite *convergente* si la suite $\{\varphi^j\}$ tend vers la solution exacte φ^* du système (1.1) quel que soit le vecteur initial, et elle est *divergente* dans le cas contraire. Une condition nécessaire et suffisante de convergence de (1.2) est évidemment que les suites $\{\psi^j\}$ et $\{\xi^j\}$ tendent vers un vecteur nul pour tous les ψ^0 et ξ^0 ($A\psi^0 = \xi^0$).

La méthode itérative (1.2) s'appelle *stationnaire* si la matrice H_j ne dépend pas du numéro de l'itération (l'opérateur T_j constitue une matrice à éléments constants), et elle est *non stationnaire* dans le cas contraire. Les méthodes itératives cycliques relèvent en principe des deux types.

Une méthode itérative est *cyclique* si elle présente la propriété : « $H_j = H_{j+s}$ pour tout $j \geqslant 0$ et un $s \geqslant 1$ fixe ». En réunissant les s itérations successives au sein d'un cycle, on aboutit à un processus stationnaire de la forme

$$\varphi^{j+1} = \varphi^j - \widetilde{H}(A\varphi^j - f), \tag{1.7}$$

avec $\widetilde{H}$ définie par l'équation

$$E - \widetilde{H}A = \prod_{i=0}^{s-1} (E - H_i A). \tag{1.8}$$

D'autre part, les méthodes cycliques sont par définition non stationnaires.

Un des problèmes essentiels en matière des méthodes itératives consiste à les optimiser, i.e. on choisit une suite des matrices H_j de classe donnée de façon à augmenter l'efficacité de l'algorithme. Le rôle de fonction objectif à minimiser revient en l'occurrence au nombre total d'opérations logiques et arithmétiques à effectuer pour trouver une solution exacte à ε près ($\varepsilon > 0$ donné). On ne résout complètement ce problème que dans des cas rarissimes, si bien qu'on minimise en réalité une fonction

$$W(\{H_j\}, \varepsilon), \tag{1.9}$$

qui majore la fonction objectif et l'approche assez bien dans un sens. Si l'on suppose qu'une itération (1.2) demande un même nombre W_0 d'opérations logiques et arithmétiques quelle que soit H_j de classe donnée, il est raisonnable de définir W par la relation

$$W = W_0 \times N(\{H_j\}, \varepsilon), \tag{1.10}$$

$N(\{H_j\}, \varepsilon)$ étant le nombre d'itérations de la méthode (1.2) au bout desquelles une norme de l'erreur initiale ψ^0 diminue de $\frac{1}{\varepsilon}$ fois. Autrement dit, $N(\{H_j\}, \varepsilon)$ est égal au nombre d'itérations (1.2) qui donnent une solution exacte à ε près du système (1.1) si la norme du vecteur erreur ψ^0 vaut 1. Le problème de définition de la fonctionnelle N et la recherche de son minimum seront illustrés sur des exemples concrets des numéros suivants.

4.2. Certaines méthodes itératives et leur optimisation

4.2.1. Méthode intérative simple

Soit le système

$$A\varphi = f \tag{2.1}$$

à matrice A définie positive symétrique dont on connaît les bornes du spectre : $\beta = \beta(A) \geqslant \alpha = \alpha(A) > 0$. On applique la méthode itérative

$$\varphi^{j+1} = \varphi^j - \tau(A\varphi^j - f) \tag{2.2}$$

avec τ réel et φ^0 un vecteur initial. On a pour les vecteurs erreurs $\psi^j = \varphi^j - \varphi^*$:

$$\psi^{j+1} = T_\tau \psi^j, \tag{2.3}$$

où

$$T_\tau = E - \tau A \tag{2.4}$$

est l'opérateur de transition.

Désignons par $\{\lambda_n\}$ l'ensemble des valeurs propres de A et par $\{u_n\}$ le système orthonormé complet de vecteurs propres associés. Développons $\{\psi^j\}$ suivant la base orthonormée $\{u_n\}$:

$$\psi^j = \sum_n \psi_n^j u_n, \qquad \psi_n^j = (\psi^j, u_n)$$

et utilisons (2.3), il vient

$$\psi_n^{j+1} = (1 - \tau\lambda_n)\,\psi_n^j.$$

Il en résulte que les coefficients ψ^j convergent vers 0 pour $j \to \infty$ et tout ψ^0 si et seulement si

$$| 1 - \tau\lambda_n | < 1,$$

et que la condition

$$q(\tau) = \beta(T_\tau) = \max_n | 1 - \tau\lambda_n | < 1 \tag{2.5}$$

est suffisante et nécessaire pour que la norme euclidienne de ψ^j

$$\| \psi^j \|_2 = (\psi^j, \psi^j)^{1/2} = \left[\sum_n (\psi_n^j)^2\right]^{1/2}$$

tend vers 0.

Voyons de plus près la dernière condition. Le système

$$-1 < 1 - \tau\lambda_n < 1, \qquad \lambda_n \in [\alpha, \beta],$$

équivalent à l'inégalité (2.5) nous fait conclure que (2.5) n'est juste que si

$$0 < \tau < \min_n \frac{2}{\lambda_n} = \frac{2}{\beta}, \tag{2.6}$$

i.e. pour $\tau \in \left(0, \frac{2}{\beta}\right)$. Ainsi, on définit le domaine de variation de τ qui garantit la convergence de la méthode (2.2) connaissant, ou bien β, ou bien sa majoration.

Comment optimiser la méthode décrite ? Considérons l'inégalité

$$\| \psi^j \|_2 \leqslant \| T_\tau^j \|_2 \, \| \psi^0 \|_2, \tag{2.7}$$

où

$$\| T_\tau^j \|_2 = [\beta(T_\tau)]^j = [q(\tau)]^j \tag{2.8}$$

par suite de la symétrie de T_τ. On voit sans peine qu'on diminue $\| \psi^0 \|_2$ de $\frac{1}{\varepsilon}$ fois en N itérations, N étant défini moyennant

$$\| T_\tau^N \|_2 = \varepsilon. \tag{2.9}$$

D'où, compte tenu de (2.8),

$$N = \left[\frac{\ln \varepsilon}{\ln [q(\tau)]}\right] + 1. \tag{2.10}$$

Comme le nombre W_0 d'opérations logiques et arithmétiques par itération ne varie pas avec τ, la fonctionnelle W de (1.10) se définit dans ce cas par la relation

$$W = W_0 \left(\left[\frac{\ln \varepsilon}{\ln [q(\tau)]} \right] + 1 \right). \tag{2.11}$$

(Ici on suppose $\tau \in \left(0, \frac{2}{\beta}\right)$.) Le problème d'optimisation (la minimisation de W en τ) se ramène, on le vérifie facilement, à minimiser $q(\tau)$ par rapport à $\tau \in \left(0, \frac{1}{\beta}\right)$. Le schéma de résolution est le suivant.

On trouve de suite que

$$q(\tau) = \max \{| 1 - \tau\alpha |, | 1 - \tau\beta |\} \tag{2.12}$$

et que la valeur optimale de τ est solution de l'équation

$$1 - \tau_{\text{opt}}\, \alpha = -(1 - \tau_{\text{opt}}\, \beta)$$

et est calculée comme suit :

$$\tau_{\text{opt}} = \frac{2}{\beta + \alpha}. \tag{2.13}$$

Le report dans (2.12) donne

$$q_{\text{opt}} = q(\tau_{\text{opt}}) = \frac{\beta - \alpha}{\beta + \alpha} = \frac{p-1}{p+1}, \tag{2.14}$$

où

$$p \equiv p(A) = \frac{\beta}{\alpha} \tag{2.15}$$

constitue le *nombre de conditionnement* de la matrice A définie positive symétrique.

Introduisons la quantité

$$R(T_\tau) = -\ln q(\tau), \tag{2.16}$$

vitesse de convergence asymptotique de la méthode (2.2). Evidemment $R^{-1}(T_\tau)$ est égal au nombre d'itérations suffisantes et nécessaires (pour ψ^0 quelconque) pour que $\| \psi^0 \|_2$ diminue de e fois, $e = 2, 7, \ldots$, étant la base des logarithmes népériens. La vitesse de convergence asymptotique constitue un bon critère de comparaison des méthodes dont on ne connaît pas le nombre d'opérations logiques et arithmétiques par itération. La relation (2.11) se récrit

$$W = W_0 \left(\left[\frac{|\ln \varepsilon|}{R(T_\tau)} \right] + 1 \right). \tag{2.17}$$

Disons quelques mots au sujet des *matrices mal conditionnées* ($p \gg 1$). Lorsque $p \gg 1$, on a

$$q_{\text{opt}} \approx 1 - \frac{2}{p}, \tag{2.18}$$

si bien que

$$R(T_\tau) \approx \frac{2}{p}, \tag{2.19}$$

donc

$$W \approx W_0 \mid \ln \varepsilon \mid \frac{p}{2}. \tag{2.20}$$

Cette analyse est tout indiquée pour comparer la qualité des méthodes à matrices mal conditionnées.

4.2.2. Convergence et optimisation des méthodes itératives stationnaires

Il est déjà apparu qu'une condition nécessaire et suffisante de convergence de (2.2) avec φ^0 arbitraire est l'inégalité

$$\beta(T) = \max_n \mid \lambda_n(T) \mid < 1, \tag{2.21}$$

où $\lambda_n(T)$ sont les valeurs propres de l'opérateur de transition T. Nous allons démontrer que c'en est également une pour la méthode stationnaire

$$\varphi^{j+1} = \varphi^j - H(A\varphi^j - f),$$

où l'opérateur de transition

$$T = E - HA$$

est une matrice constante.

Soit

$$J = \begin{Vmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & J_l \end{Vmatrix}$$

la forme normale de Jordan de la matrice T (i.e. $T = SJS^{-1}$, les colonnes de S étant formées de vecteurs propres et de vecteurs principaux de T) avec les blocs de Jordan

$$J_i = \begin{Vmatrix} \lambda_i & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_i & 1 & \dots & 0 & 0 \\ 0 & 0 & \lambda_i & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_i & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda_i \end{Vmatrix}$$

d'ordre $k_i \geqslant 1$ associés aux valeurs propres $\lambda_i = \lambda_i(T)$, $i = 1, \ldots, l$. Du moment que

$$\psi^j = T^j \psi^0 = S J^j S^{-1} \psi^0,$$

$\{\psi^j\}$ étant les vecteurs erreurs, on vérifie immédiatement que pour qu'il y ait convergence de la méthode, il faut et il suffit que les matrices T^j (et, partant, J^j) aient pour limite une matrice nulle quand $j \to \infty$. On n'a ce dernier résultat que si J_i^j convergent pour $j \to \infty$ vers une matrice nulle quel que soit i, $1 \leqslant i \leqslant l$.

Ainsi, on a l'affirmation voulue sous la condition suffisante que J_i^j tendent vers des matrices nulles, quand $j \to \infty$, seulement pour $|\lambda_i| < 1$. Désignons par $\alpha_{s,t}^{(j)}$ les éléments de J_i^j. On montre par multiplication directe que

$$\alpha_{s,t}^{(j)} = \left\{ \begin{array}{ll} 0 & \text{si } s > t, \\ \lambda_i^j & \text{si } s = t, \\ C_j^{t-s} \lambda_i^{j-t+s} & \text{si } s < t, \end{array} \right\} (j \geqslant k_i - 1),$$

où $C_j^{t-s} = \dfrac{j(j-1)\ldots(j-t+s+1)}{(t-s)!}$ sont des coefficients binomiaux. Autrement dit,

$$J_i^j = \begin{Vmatrix} \lambda_i^j & C_j^1 \lambda^{j-1} & \ldots & C_j^{k_i-1} \lambda_i^{j-k_i+1} \\ 0 & \lambda_i^j & \ldots & C_j^{k_i-2} \lambda_i^{j-k_i+2} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \lambda_i^j \end{Vmatrix}$$

quels que soient $j \geqslant k_i - 1$ et $1 \leqslant i \leqslant l$. On montre par des calculs élémentaires que $\alpha_{s,t}^{(j)}$ tendent vers 0 (resp. J_i^j convergent vers une matrice nulle) pour $j \to \infty$ et n'importe quels $1 \leqslant s, t \leqslant k_i$ seulement si $|\lambda_i| < 1$, c.q.f.d. Il est connu que ce résultat reste en vigueur pour toute méthode itérative stationnaire (voir p.ex. Faddeev, Faddeeva [VIII]).

Dans ce point nous aborderons l'optimisation des méthodes itératives stationnaires avec une optique générale, à l'aide de leur comportement asymptotique. Nous aurons besoin d'un résultat connu de l'analyse fonctionnelle (voir Kantorovich, Akilov [I]) qui s'énonce comme suit pour les matrices.

Toute matrice carrée T et toute norme matricielle $\|\cdot\|$ vérifient l'égalité

$$\lim_{k \to \infty} \|T^k\|^{1/k} = \beta(T). \tag{2.22}$$

Supposons que la matrice H_τ de la méthode itérative stationnaire

$$\varphi^{j+1} = \varphi^j - H_\tau (A\varphi^j - f) \tag{2.23}$$

dépend des paramètres $\tau_1, \tau_2, \ldots, \tau_s$. Si le nombre W_0 d'opérations logiques et arithmétiques par itération reste le même avec τ variant, alors l'optimisation de la méthode (2.23) consiste (voir n° 4.1) à minimiser la fonction

$$W = W_0 \times N(H_\tau, \varepsilon) \tag{2.24}$$

par rapport à $\tau_1, \ldots, \tau_s$, $N(H_\tau, \varepsilon)$ étant le nombre d'itérations au bout desquelles une norme du vecteur erreur ψ^0 diminue de $\frac{1}{\varepsilon}$ fois.

Comme

$$\|\psi^j\| \leqslant \|T_\tau^j\| \, \|\psi^0\|, \tag{2.25}$$

avec $T_\tau = E - H_\tau A$ et $\|\cdot\|$ une norme, N dépend de H_τ et ε selon la formule

$$\|T_\tau^N\| = \varepsilon. \tag{2.26}$$

La solution exacte de cette équation n'est évidemment obtenue que dans des cas fort rares (voir point 4.2.1). Aussi on en cherche une approximation sous des hypothèses complémentaires qui varient selon le cas concerné.

On établit approximativement la dépendance mentionnée entre N, d'une part, et H_τ et ε de l'autre au sens asymptotique, à savoir on suppose que $\varepsilon \ll 1$ et $N \gg 1$. Le théorème ci-dessus entraîne

$$\|T_\tau^N\| = [\|T_\tau^N\|^{1/N}]^N \approx [\beta(T_\tau)]^N,$$

$$N(H_\tau, \varepsilon) \approx \frac{|\ln \varepsilon|}{R(T_\tau)}, \tag{2.27}$$

$R(T_\tau) = -\ln \beta(T_\tau)$ étant la vitesse de convergence asymptotique de la méthode (2.23). Ainsi, la fonction W se définit par

$$W = W_0 \frac{|\ln \varepsilon|}{R(T_\tau)}. \tag{2.28}$$

On note que (2.17) est un cas particulier de (2.28).

Conformément à la dernière formule, le problème d'optimisation de (2.23) consiste avec les hypothèses faites à maximiser $R(T_\tau)$ (à minimiser $\beta(T_\tau)$) par rapport à $\tau_1, \ldots, \tau_s$ tels que $\beta(T_\tau) < 1$.

Quelques mots enfin sur les méthodes itératives cycliques avec $s \geqslant 1$. Une méthode cyclique se ramène, on l'a montré au n° 4.1, à la méthode stationnaire ordinaire (1.7) avec l'opérateur de transition

$$\widetilde{T}_\tau = \prod_{i=0}^{s-1} (E - H_i A). \tag{2.29}$$

Admettons que chaque itération de (1.7) demande sW_0 opérations logiques et arithmétiques et qu'un pas de la méthode cyclique en exige W_0. On a

$$W = W_0 \frac{|\ln \varepsilon|}{R(\widetilde{T}_\tau)}, \tag{2.30}$$

où $R(\widetilde{T}_\tau) = -\frac{1}{s} \ln \beta(\widetilde{T}_\tau)$. S'agissant des méthodes cycliques, l'optimisation consiste donc à chercher le minimum de $\beta(\widetilde{T}_\tau)$ pour s fixe.

4.2.3. Surrelaxation successive

La méthode de surrelaxation successive élaborée par Young [X] et Frankel [X] s'avère très utile dans beaucoup d'applications. Voici son idée directrice mise en œuvre dans le cas du système algébrique linéaire

$$A\varphi = f \tag{2.31}$$

à matrice tripldiagonale par blocs

$$A = \begin{Vmatrix} E_1 & -S_1 & \dots & 0 & 0 \\ -R_2 & E_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & E_{k-1} & -S_{k-1} \\ 0 & 0 & \dots & -R_k & E_k \end{Vmatrix}, \tag{2.32}$$

où E_l sont des matrices unités d'ordre n_l et R_l, S_l des matrices d'ordre respectif $n_l \times n_{l-1}$ et $n_l \times n_{l+1}$. Les systèmes à de telles matrices interviennent souvent lorsqu'on réduit les équations elliptiques à des systèmes d'équations aux différences finies ou d'équations variationnelles aux différences.

Mettons la matrice A sous la forme

$$A = E - R - S, \tag{2.33}$$

où

$$R = \begin{Vmatrix} 0 & 0 & \dots & 0 & 0 \\ R_2 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & R_k & 0 \end{Vmatrix}, \quad S = \begin{Vmatrix} 0 & S_1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & S_{k-1} \\ 0 & 0 & \dots & 0 & 0 \end{Vmatrix},$$

E étant une matrice unité. Si l'on introduit la matrice

$$B = R + S, \tag{2.34}$$

le système (2.31) s'écrit

$$\varphi = B\varphi + f. \tag{2.35}$$

On supposera partout dans la suite que la méthode itérative

$$\varphi^{j+1} = B\varphi^j + f \tag{2.36}$$

converge (i.e. $\beta(B) < 1$), que la matrice B a toutes ses valeurs propres réelles et qu'elle possède un système complet de vecteurs propres.

Mettons (2.31) sous forme de système d'équations matricielles

$$\Phi_l - R_l\Phi_{l-1} - S_l\Phi_{l+1} = F_l,\ l = 1, \ldots, k, \tag{2.37}$$

avec les conditions

$$R_1 = S_k = 0 \text{ et } \Phi_0 = \Phi_{k+1} = 0,$$

où Φ_l et F_l sont les composantes vectorielles des vecteurs φ et f. La méthode de surrelaxation successive est alors définie par les formules

$$\Phi_l^{j+1} = \Phi_l^j - \tau(\Phi_l^j - R_l\Phi_{l-1}^{j+1} - S_l\Phi_{l+1}^j - F_l),\quad l = 1, \ldots, k, \tag{2.38}$$

avec τ un paramètre réel. Ces relations se récrivent

$$\Phi_l^{j+1/2} = R_l\Phi_{l-1}^{j+1} + S_l\Phi_{l+1}^j + F_l,$$
$$\Phi_l^{j+1} = \tau\Phi_l^{j+1/2} + (1-\tau)\Phi_l^j,$$
$$l = 1, \ldots, k.$$

On voit que le calcul se fait pour chaque j bloc par bloc, Φ_0^{j+1} et Φ_{k+1}^j étant supposés nuls.

Compte tenu de (2.38), on établit les équations des vecteurs erreurs $\Psi_l^j = \Phi_l^j - \Phi_l^*$, où $\{\Phi_l^*\}$ est solution du système (2.37):

$$\Psi_l^{j+1} = \Psi_l^j - \tau(\Psi_l^j - R_l\Psi_{l-1}^{j+1} - S_l\Psi_{l+1}^j),\qquad l = 1, \ldots, k. \tag{2.39}$$

On en tire que l'opérateur de transition de la méthode (2.37) est défini par l'égalité

$$T_\tau = (E - \tau R)^{-1}[(1-\tau)E + \tau S] \tag{2.40}$$

et dépend d'un seul τ et que le problème spectral

$$T_\tau\Psi = \lambda(T_\tau)\Psi \tag{2.41}$$

se transforme sans peine en

$$\lambda(T_\tau)\Psi_l = \Psi_l - \tau[\Psi_l - \lambda(T_\tau)R_l\Psi_{l-1} - S_l\Psi_{l+1}],$$
$$l = 1, \ldots, k, \tag{2.42}$$

dont la solution est cherchée sous la forme

$$\Psi_l = [\lambda(T_\tau)]^{1/2}w_l, \tag{2.43}$$

$\{w_l\}_{l=1}^{k}$ étant les composantes vectorielles du vecteur propre w du problème spectral

$$Bw = \lambda(B)\, w \tag{2.44}$$

et $w_0 = w_{k+1} = 0$. Portons (2.43) dans (2.42), il vient par suite de (2.44)

$$\lambda^{l/2}(T_\tau)\,[\lambda(T_\lambda) - \tau\lambda^{1/2}(T_\tau)\,\lambda(B) + \tau - 1]\, w_l = 0,$$
$$l = 1, \ldots, k. \tag{2.45}$$

Les valeurs nulles $\lambda(T_\tau)$ n'influent pas sur la convergence de la méthode étudiée, et parmi $\{w_l\}$ il se trouve au moins un vecteur non nul, si bien que (2.45) conduit à

$$\lambda(T_\tau) - \tau\lambda^{1/2}(T_\tau)\,\lambda(B) + \tau - 1 = 0,$$

relation entre les valeurs propres des matrices T_τ et B. Sa résolution donne

$$\lambda^{1/2}(T_\tau) = \frac{\tau\lambda(B)}{2} \pm \sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1}. \tag{2.46}$$

On observe que si $\lambda(B)$ est valeur propre de B, il en est également de $-\lambda(B)$. En effet, si $\lambda(B)\, w = Bw$, où w est un vecteur propre de composantes vectorielles $\{w_l\}_{l=1}^{k}$ de la matrice B, alors $\widetilde{w}$ de composantes vectorielles $\widetilde{w}_l = (-1)^l w_l$ est un vecteur propre de B associé à $-\lambda(B)$, i.e.

$$-\lambda(B)\,\widetilde{w} = B\widetilde{w}. \tag{2.47}$$

Avec ce résultat et la formule (2.46), on se borne, en analysant $\lambda^{1/2}(T_\tau)$, au cas $\lambda(B) \geqslant 0$.

Considérons $\lambda^{1/2}(T_\tau)$ comme étant une fonction du paramètre τ et admettons que $\lambda(B) < 1$ est une valeur propre non négative fixe de B. Soit $\tau \leqslant 0$. On établit sans peine l'impossibilité de

$$|\lambda^{1/2}(T_\tau)| = \left|\frac{\tau\lambda(B)}{2} \pm \sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1}\right| < 1, \tag{2.48}$$

si bien que la positivité de τ constitue une condition nécessaire de convergence de la surrelaxation successive.

Si l'on se borne à τ positif et si l'on se rappelle que la convergence de la méthode est définie par la seule valeur propre de plus grand module de T_τ, il suffit, on constate sans peine, d'étudier la formule

$$\lambda^{1/2}(T_\tau) = \frac{\tau\lambda(B)}{2} + \sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1},$$

où la racine carrée est prise positive dès qu'il en est de l'expression $\frac{\tau^2\lambda^2(B)}{4} - \tau + 1$.

Si $\tau \in [\tau_1, \tau_2]$, avec

$$\tau_{1,2} = \frac{2}{\lambda^2(B)}\,[1 \pm \sqrt{1 - \lambda^2(B)}],$$

$\frac{\tau^2\lambda^2(B)}{4} - \tau + 1$ de (2.48) est non positive et on a pour ces τ:

$$|\lambda(T_\tau)| = |\lambda^{1/2}(T_\tau)|^2 = \left(\frac{\tau\lambda(B)}{2}\right)^2 - \left(\frac{\tau^2\lambda^2(B)}{4} - \tau + 1\right) = \tau - 1.$$

Lorsque $\tau \notin [\tau_1, \tau_2]$, l'expression sous le radical dans (2.48) est toujours positive et $\lambda^{1/2}(T_\tau)$ est non négative. Ceci étant, on montre sans difficulté que

$$\lambda^{1/2}(T_\tau) \geqslant 1$$

ne possède pas de solution pour $\tau \in (0, \tau_1)$ et qu'on a

$$\lambda^{1/2}(T_\tau) \geqslant \frac{\tau\lambda(B)}{2} \geqslant \frac{\tau_2\lambda(B)}{2} = \frac{1}{\lambda(B)}[1 + \sqrt{1-\lambda^2(B)}] > 1$$

quand $\tau \geqslant \tau_2$. Ces résultats autorisent à dire qu'avec les hypothèses faites,

$$\tau \in (0, 2)$$

est une condition nécessaire et suffisante pour que la méthode de surrelaxation successive (2.38) soit convergente.

Etudions maintenant

$$\lambda^{1/2}(T_\tau) = \begin{cases} \frac{\tau\lambda(B)}{2} + \sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1}, & \tau \leqslant \tau_1, \\ \sqrt{\tau - 1}, & \tau \geqslant \tau_1, \end{cases}$$

comme fonction du paramètre τ. Puisque

$$\frac{\partial}{\partial\tau}[\lambda^{1/2}(T_\tau)] = \frac{\lambda(B)\lambda^{1/2}(T_\tau) - 1}{2\sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1}} < 0$$

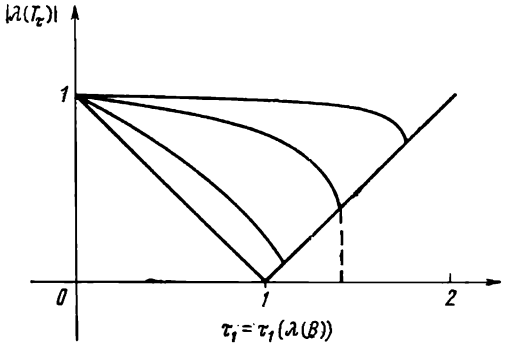


Fig. 13

pour $\tau \in (0, \tau_1]$ et $\lambda^{1/2}(T_\tau) = 1$ si $\tau = 0$, $|\lambda(T_\tau)|$ est représenté pour $\lambda(B)$ différentes par les courbes de la fig. 13.

Comme

$$\frac{\partial}{\partial\lambda(B)}[\lambda^{1/2}(T_\tau)]=\begin{cases}\dfrac{\tau\lambda^{1/2}(T_\tau)}{2\sqrt{\dfrac{\tau^2\lambda^2(B)}{4}-\tau+1}}>0, & \tau<\tau_1,\\ 0, & \tau\geqslant\tau_1,\end{cases}$$

on en tire pour τ de l'intervalle (0, 2)

$$\beta(T_\tau)=\begin{cases}\dfrac{\tau\beta(B)}{2}+\sqrt{\dfrac{\tau^2\beta^2(B)}{4}-\tau+1}, & \tau<\tau_1,\\ \tau-1, & \tau\geqslant\tau_1.\end{cases}$$

Cette formule entraîne de suite que la valeur minimale de $\beta(T_\tau)$ est atteinte pour $\tau=\tau_1$. Ainsi, τ_{opt} maximisant la vitesse de convergence asymptotique $R(T_\tau)$ de la méthode étudiée est calculé par la formule

$$\tau_{opt}=\frac{2}{\beta^2(B)}[1-\sqrt{1-\beta^2(B)}]=\frac{2}{1+\sqrt{1-\beta^2(B)}}, \tag{2.49}$$

et

$$R(T_{\tau_{opt}})=-\ln(\tau_{opt}-1)=-\ln\frac{1-\sqrt{1-\beta^2(B)}}{1+\sqrt{1-\beta^2(B)}}. \tag{2.50}$$

Voyons ce que deviennent les formules (2.49), (2.50) pour les matrices mal conditionnées (et A symétrique), i.e. $p(A)=\dfrac{\beta(A)}{\alpha(A)}\gg$ $\gg 1$. On sait déjà que si B de la forme (2.34) possède une valeur propre $\beta(B)$, elle admet une valeur propre $-\beta(B)$, si bien que la relation $B=E-A$ entraîne

$$\beta(B)=1-\alpha(A),$$

et

$$-\beta(B)=1-\beta(A),$$

d'où $\alpha(A)=2-\beta(A)$ et la condition $p(A)\gg 1$ signifie que

$$\alpha(A)\approx\frac{2}{p},$$

$$\beta(B)\approx 1-\frac{2}{p}.$$

Portons dans (2.49), (2.50), il vient asymptotiquement

$$\tau_{opt}\approx 2-\frac{4}{\sqrt{p}},\qquad \beta(T_{\tau_{opt}})\approx 1-\frac{4}{\sqrt{p}},$$
$$R(T_{\tau_{opt}})\approx\frac{4}{\sqrt{p}},\qquad W\approx W_0\frac{|\ln\varepsilon|}{4}\sqrt{p}. \tag{2.51}$$

S'agissant des matrices mal conditionnées, la vitesse de convergence asymptotique de la surrelaxation successive avec paramètres optimaux est donc $2\sqrt{p}$ fois celle de la méthode (2.2). On observe qu'à toute itération concrète (surtout au début du processus), la vitesse de convergence est inférieure à celle de (2.51). La cause en est la présence d'un bloc d'ordre 2 dans $T_{\tau_{opt}}$ sous forme de Jordan (voir Young [X], Tee [X]).

4.2.4. Méthode itérative de Tchebycheff

Soit le système d'équations algébriques linéaires

$$A\varphi = f \tag{2.52}$$

à matrice A définie positive symétrique. On le résout itérativement par les formules

$$\varphi^{j+1} = \varphi^j - \tau_j (A\varphi^j - f), \tag{2.53}$$

connues sous le nom de *méthode de Richardson*. On l'optimise en choisissant une suite des paramètres τ_j telle que φ^j converge au plus vite vers la solution exacte du système (2.52).

La méthode cyclique (2.53) où $s \geqslant 1$ s'interprète comme un procédé itératif stationnaire

$$\varphi^{j+\frac{i}{s}} = \varphi^{j+\frac{i-1}{s}} - \tau_i (A\varphi^{j-\frac{i-1}{s}} - f), \qquad i = 1, \ldots, s, \tag{2.54}$$

à opérateur de transition

$$\widetilde{T}_\tau = \prod_{i=1}^{s} (E - \tau_i A). \tag{2.55}$$

La symétrie de A fait que

$$\beta(\widetilde{T}_\tau) = \|\widetilde{T}_\tau\|_2 = \max_n \left| \prod_{i=1}^{s} (1 - \tau_i \lambda_n(A)) \right| \tag{2.56}$$

et l'optimisation de (2.53) consiste, selon le point 4.2.2, à minimiser $\beta(\widetilde{T}_\tau)$ par rapport aux paramètres $\tau_1, \ldots, \tau_s$. La solution exacte pour $s = 1$ est connue dès le point 4.2.1. Si $s > 1$ le problème n'admet pas de solution exacte parce qu'il ne suffit plus de connaître les bornes du spectre de A. Or, la recherche de toutes les valeurs propres de cette matrice est une tâche plus complexe que la résolution de (2.52) par la méthode itérative simple. L'optimisation exige donc qu'on minimise la fonction

$$q_s(\tau) = \max_{\alpha \leqslant \lambda \leqslant \beta} \left| \prod_{i=1}^{s} (1 - \tau_i \lambda) \right|, \tag{2.57}$$

$0 < \alpha = \alpha(A) \leqslant \beta = \beta(A)$, qui majore $\beta(\widetilde{T}_\tau)$ et l'approche suffisamment bien dans la pratique.

On se propose, en plus de la recherche du minimum de $q_s(\tau)$, de former un polynôme $\widetilde{P}_s(\lambda)$ de degré s, solution du problème

$$\min_{P_s(\lambda)\in Q_s} \max_{\alpha\leqslant\lambda\leqslant\beta} |P_s(\lambda)| = \max_{\alpha\leqslant\lambda\leqslant\beta} |\widetilde{P}_s(\lambda)|, \tag{2.58}$$

Q_s étant l'ensemble des $P_s(\lambda)$ de degré s vérifiant la condition $P_s(0) = 1$. Markov a résolu ce problème à l'aide des polynômes de Tchebycheff :

$$\widetilde{P}_s(\lambda) = \frac{T_s\left(\frac{\beta+\alpha-2\lambda}{\beta-\alpha}\right)}{T_s\left(\frac{\beta+\alpha}{\beta-\alpha}\right)}, \tag{2.59}$$

où

$$T_s(t) = \frac{(t+\sqrt{t^2-1})^s + (t-\sqrt{t^2-1})^s}{2} \tag{2.60}$$

est le polynôme de Tchebycheff de degré s. La propriété fondamentale de $\frac{1}{2^{s-1}} T_s(t)$ est que parmi tous les polynômes de degré s et de coefficient de t^s égal à 1, il s'écarte en norme uniforme le moins possible de zéro sur le segment $[-1, 1]$. Les racines de $T_s(t)$ s'obtiennent par la formule

$$x_i = \cos\frac{2i-1}{2s}\pi, \qquad i = 1, \ldots, s,$$

ses valeurs de plus grand module sont atteintes aux points

$$y_i = \cos\frac{i\pi}{2}, \qquad i = 0, 1, \ldots, s,$$

et

$$T_s(y_i) = (-1)^i, \qquad i = 0, 1, \ldots, s.$$

$\{\lambda_i\}$ définies à partir de

$$x_i = \frac{\beta+\alpha-2\lambda_i}{\beta-\alpha}, \qquad i = 1, \ldots, s,$$

constituent donc les racines de $T_s\left(\frac{\beta+\alpha-2\lambda}{\beta-\alpha}\right)$, si bien que

$$T_s\left(\frac{\beta+\alpha-2\lambda}{\beta-\lambda}\right) = r_s \prod_{i=1}^{s} (\lambda_i - \lambda),$$

avec r_s une constante. *D'où*

$$\widetilde{P}_s(\lambda) = \frac{\prod_{i=1}^{s}(\lambda_i - \lambda)}{\prod_{i=1}^{s}\lambda_i} = \prod_{i=1}^{s}\left(1 - \frac{1}{\lambda_i}\lambda\right), \tag{2.61}$$

et

$$\max_{\alpha\leqslant\lambda\leqslant\beta}|\widetilde{P}_s(\lambda)| = \frac{1}{T_s\left(\frac{\beta+\alpha}{\beta-\alpha}\right)} \max_{\alpha\leqslant\lambda\leqslant\beta}\left|T_s\left(\frac{\beta+\alpha-2\lambda}{\beta-\alpha}\right)\right| =$$

$$= \frac{1}{T_s\left(\frac{\beta+\alpha}{\beta-\alpha}\right)} \max_{-1\leqslant t\leqslant 1}|T_s(t)| = \frac{1}{T_s\left(\frac{\beta+\alpha}{\beta-\alpha}\right)}. \tag{2.62}$$

Revenons au problème d'optimisation de (2.53), i.e. au problème d'extrémum

$$q_s(\tau_{\text{opt}}) = \min_{\tau} q_s(\tau), \tag{2.63}$$

τ étant une suite de paramètres $\tau_1, \ldots, \tau_s$. En notant V_s l'ensemble des polynômes de degré s

$$P_s(\lambda) = \prod_{i=1}^{s}(1 - a_i\lambda), \tag{2.64}$$

on énonce le problème (2.63) comme suit : trouver $\hat{P}_s(\lambda)$ solution du problème

$$\min_{P_s(\lambda)\in V_s}\max_{\alpha\leqslant\lambda\leqslant\beta}|P_s(\lambda)| = \max_{\alpha\leqslant\lambda\leqslant\beta}|\hat{P}(\lambda)|. \tag{2.65}$$

Comme $V_s \subset Q_s$, on a évidemment

$$\max_{\alpha\leqslant\lambda\leqslant\beta}|\widetilde{P}_s(\lambda)| \leqslant \max_{\alpha\leqslant\lambda\leqslant\beta}|\hat{P}_s(\lambda)|. \tag{2.66}$$

Le problème (2.58) admet pour solution un polynôme $\widetilde{P}_s(\lambda)$ de la forme (2.64), i.e. un élément de V_s, si bien que (2.66) devient une égalité stricte. Cela permet de dire que le rôle de $\hat{P}_s(\lambda)$ peut incomber à $\widetilde{P}_s(\lambda)$. Finalement, le choix

$$\hat{P}_s(\lambda) = \widetilde{P}_s(\lambda) \tag{2.67}$$

est unique vu l'unicité de la solution de (2.58) *).

*) Voir p.ex. Flanders, Shortley [VIII].

Si l'on désigne par $\varkappa_s(\sigma_1, \sigma_2, \ldots, \sigma_s)$, $1 \leqslant \sigma_i \leqslant s$, $\sigma_i \neq \sigma_k$, $i \neq k$, $i, k = 1, 2, \ldots, s$, une permutation d'ordre s, alors (2.62) entraîne pour la méthode itérative optimale (2.54)

$$\tau_i = \frac{1}{\lambda_{\sigma_i}}, \qquad i = 1, \ldots, s, \tag{2.68}$$

où

$$\lambda_i = \frac{1}{2}[\beta + \alpha - (\beta - \alpha) x_i], \qquad i = 1, \ldots, s.$$

Nous appellerons *méthode itérative cyclique de Tchebycheff* le procédé (2.54) avec paramètres définis par (2.68).

On se propose d'en évaluer la convergence.

On montre que s peut être choisi par la condition « s itérations de (2.53) (i.e. une itération de (2.54)) garantissent une erreur initiale $1/\varepsilon$ fois moindre ». On considère l'équation

$$\varepsilon = \frac{1}{T_s\left(\frac{\beta+\alpha}{\beta-\alpha}\right)}.$$

Avec les notations

$$t_0 = \frac{\beta+\alpha}{\beta-\alpha}, \qquad \gamma = t_0 - \sqrt{t_0^2 - 1}$$

et la relation

$$t_0 - \sqrt{t_0^2-1} = \frac{1}{t_0 + \sqrt{t_0^2-1}},$$

cette équation devient par de simples transformations :

$$\varepsilon = \frac{2\gamma^s}{\gamma^{2s}+1}.$$

Cette relation regardée comme une équation du deuxième degré en γ^s donne, compte tenu de $\gamma < 1$,

$$\gamma^s = \frac{1-\sqrt{1-\varepsilon^2}}{\varepsilon} = \frac{\varepsilon}{1+\sqrt{1-\varepsilon^2}},$$

donc

$$s = \frac{\ln \frac{\varepsilon}{1+\sqrt{1-\varepsilon^2}}}{\ln \gamma}.$$

Lorsque $\varepsilon \ll 1$ et $p(A) \gg 1$, on a

$$s \approx \frac{|\ln \varepsilon|\sqrt{p}}{2}.$$

d'où

$$R(\tilde{T}_{\tau_{opt}}) \approx \frac{2}{\sqrt{p}},$$

$$W \approx W_0 |\ln \varepsilon| \frac{\sqrt{p}}{2}. \tag{2.69}$$

Il en résulte que la méthode de Tchebycheff converge asymptotiquement $\sqrt{p}$ fois plus vite que la méthode simple à deux niveaux et 2 fois plus lentement que la relaxation successive.

La mise en œuvre de la méthode de Tchebycheff exige, on le voit, qu'on connaisse les bornes α et β du spectre de A. Cela est d'autant plus nécessaire que les polynômes de Tchebycheff croissent rapidement pour $x \notin [-1, 1]$, de sorte que l'erreur sur les bornes risque de ralentir considérablement la convergence.

La borne supérieure est en général définie par le théorème de Gershgorin. La définition de la borne inférieure constitue un problème plus difficile. On utilise des fois des estimations à priori, mais dans la plupart des cas on introduit un processus itératif supplémentaire (cela peut être par exemple la méthode de Ljusternik décrite au Chapitre premier ou les itérations minimisées de Lanczos).

Un autre point délicat du procédé de Tchebycheff est l'arrangement des paramètres, car avec τ_i rangés de façon arbitraire la mise en œuvre numérique donne lieu à une instabilité. Cet inconvénient a fait couler beaucoup d'encre et a été éliminé récemment par Lébédev, Finoguénov [IX; 9] et Samarski [III; 12]. Voici deux algorithmes de numérotage des paramètres.

Soit $s = 2^r$, r étant un entier positif, et $\varkappa_{2^{r-1}} = (\sigma_1, \sigma_2, \ldots$ $\ldots, \sigma_{s/2})$ l'ordre des x_i associé à $s = 2^{r-1}$ dans l'algorithme décrit. Si $s = 2^r$, on demande que

$$\varkappa_{2^r} = (\sigma_1, s+1-\sigma_1, \sigma_2, s+1-\sigma_2, \ldots, \sigma_{s/2}, s+1-\sigma_{s/2}).$$

Disons pour $s = 16$,

$$\varkappa_{16} = (1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11).$$

Soit maintenant $s = 3^r$, r étant un entier positif, et $\varkappa_{3^{r-1}} =$ $= (\sigma_1, \sigma_2, \ldots, \sigma_{s/3})$ l'ordre des x_i pour $s = 3^{r-1}$. Lorsque $s = 3^r$, l'arrangement doit être

$$\varkappa_{3^r} = (\sigma_1, 2\cdot 3^{r-1}+\sigma_1, 2\cdot 3^{r-1}+1-\sigma_1, \ldots, 2\cdot 3^{r-1}+1-\sigma_{s/3}).$$

Le second algorithme donne par exemple pour $s = 9$:

$$\varkappa_9 = (1, 7, 6, 3, 9, 4, 2, 8, 5).$$

Avec ces algorithmes, les opérateurs de transition $T_i = E - \tau_i A$ à norme importante se répartissent de manière suffisamment uniforme parmi les opérateurs diminuant la norme de l'erreur.

La méthode (2.53) a été décrite comme étant optimale pour s donné. Lébédev et Finoguénov [IX ; 9] ont proposé une nouvelle classe de processus itératifs. Il s'agit des méthodes optimales infinies stables (2.53) avec paramètres de Tchebycheff qui permettent de poursuivre le processus (2.53) après s itérations données de façon qu'il soit stable et redevienne optimal pour certains $j = j_k$, $j_k \to \infty$, $j_k > s$. Ci-dessous l'algorithme de génération des paramètres pour un cas particulier. Comme

$$\cos 3x = \cos x\,(2\cos 2x - 1),$$

entre $T_s(t)$ et $T_{3s}(t)$ existe la relation

$$T_{3s}(t) = T_s(t)\,(2T_{2s}(t) - 1)$$

qui montre que l'ensemble des racines de $T_{3s}(t)$ comprend (2.61), racines de $T_s(t)$, et toutes les racines de $2T_{2s}(t) - 1$:

$$\widetilde{x}_i = \cos\frac{2i-1}{6s}\pi, \quad 2i \not\equiv 1 \pmod 3, \qquad 1 \leqslant i \leqslant 3s.$$

Si l'on effectue donc s itérations telles que les formules (2.68) utilisent les paramètres de (2.61) et si l'on continue avec x_i de (2.68) (λ_i s'expriment par x_i) remplacés par $\widetilde{x}_i$ ordonnés de façon convenable, alors on aboutit pour $j = 3s$ à une méthode optimale. En continuant dans cette voie, on obtient une suite infinie de x_i telle que la méthode (2.53) devienne optimale pour $j = 3^r s$. Donnons les formules de numérotage des x_i dans le cas $s = 2$ (voir Lébédev [IX ; 8]). On pose $x_1 = -2^{-1/2}$, $x_2 = -x_1$. Supposons la suite de x_i, $i = 1, \ldots, 2\cdot 3^{r-1}$, ordonnée, auquel cas la suite partielle avec $i = 2\cdot 3^{r-1} + 1, \ldots, 2\cdot 3^r$ est générée comme suit. On forme par permutation $\varkappa_{3^{r-1}}$ les quantités

$$t_{j-1} = \sin\frac{2(\sigma_j + [\sigma_{j/2}]) - 1}{4\cdot 3^r}\pi, \qquad j = 1, \ldots, 3^{r-1}.$$

Alors

$$x_{2\cdot 3^{r-1}+4j+1} = -t_j, \qquad x_{2\cdot 3^{r-1}+4j+2} = t_j,$$

$$x_{2\cdot 3^{r-1}+4j+3} = -\sqrt{1-t_j^2}, \qquad x_{2\cdot 3^{r-1}+4j+4} = \sqrt{1-t_j^2},$$

$$j = 0, \ldots, 3^{r-1} - 1.$$

On calcule x_i, $i = 2\cdot 3^r + 1, \ldots, 2\cdot 3^{r+1}$, moyennant $\varkappa_{3^r}$, et ainsi de suite.

On voit que l'optimisation de la méthode de Tchebycheff (2.53) n'a jamais utilisé la propriété (2.56), qu'on a minimisé une fonction $q_s(\tau)$ qui ne dépend que des bornes du spectre et qu'on construit compte tenu du seul caractère réel des valeurs propres de A. Ainsi, il suffit, pour optimiser la méthode, d'exiger que les valeurs propres

de A soient positives et que les vecteurs propres associés forment un système complet tout en renonçant à la symétrie de A.

A la lumière de ces faits, on construit la théorie de l'optimisation des méthodes itératives de la forme

$$\varphi^{j+1} = \varphi^j - \tau_j H (A\varphi^j - f) \tag{2.70}$$

sous l'hypothèse que toutes les valeurs propres de HA sont réelles et appartiennent au segment $[\alpha, \beta]$, où $0 < \alpha = \alpha(HA) \leqslant \beta = \beta(HA)$.

On montre que le spectre de HA vérifie toutes les conditions énumérées s'il y a symétrie et définie positivité de H et A. A cet effet, on introduit la notion de racine carrée positive d'une matrice D définie positive et symétrique. Soit $\{v_n\}$ un système orthonormé complet de vecteurs propres de D associés à ses valeurs propres positives $\{d_n\}$. Il existe une matrice orthogonale P formée de vecteurs colonnes $\{v_k\}$ telle que

$$D = PD_0P^*,$$

avec D_0 une matrice diagonale dont les éléments diagonaux sont les valeurs propres $\{d_n\}$ de D. On définit $D_0^{1/2}$ comme une matrice diagonale ayant les valeurs propres positives $\{d_n^{1/2}\}$ sur la diagonale principale et telle que

$$D_0^{1/2}D_0^{1/2} = D_0.$$

La matrice D_0 définit évidemment $D_0^{1/2}$ de façon unique. La matrice $D^{1/2}$ est donnée par l'égalité

$$D^{1/2} = PD_0^{1/2}P^*.$$

On vérifie aisément sa symétrie, sa définie positivité (vu sa symétrie et la positivité des valeurs propres $\{d_n^{1/2}\}$) et la validité des égalités

$$D^{1/2}D^{1/2} = [D^{1/2}]^2 = D.$$

On établit maintenant sans peine que HA est semblable à la matrice $S = A^{1/2}HA^{1/2}$ définie positive symétrique :

$$A^{1/2}[HA][A^{1/2}]^{-1} = A^{1/2}HA^{1/2}.$$

D'où la réalité et la positivité de toutes ses valeurs propres.

On montre que sous les hypothèses faites HA possède un système complet de vecteurs propres (ce qui permet d'utiliser, à côté des méthodes cycliques, des méthodes infinies de Tchebycheff). On désigne par $\{v_n\}$ le système orthonormé complet de vecteurs propres de la matrice S associés aux valeurs propres positives $\{\mu_n\}$:

$$Sv_n = \mu_n v_n.$$

Multiplions par $[A^{1/2}]^{-1}$, il vient avec la notation $w_n = [A^{1/2}]^{-1}v_n$:

$$HAw_n = \mu_n w_n.$$

Ainsi, $\{\mu_n\}$ sont valeurs propres de HA et $\{w_n\}$ les vecteurs propres associés. Comme le système $\{v_n\}$ forme une base dans l'espace primitif de vecteurs et $\{w_n\}$ s'obtient de $\{v_n\}$ par transformation non dégénérée, $\{w_n\}$ constitue également une base, si bien que HA possède un système complet de vecteurs propres.

Considérons un autre cas important où HA a ses valeurs propres réelles et positives. On suppose A être de la forme

$$A = \begin{Vmatrix} E_1 & -S_1 \\ -R_2 & E_2 \end{Vmatrix}, \tag{2.71}$$

E_1, E_2 étant deux matrices unité d'ordre n_1 et n_2 respectivement, et S_1 et R_2 respectivement $n_1 \times n_2$ et $n_2 \times n_1$. On suppose de plus que la matrice

$$B = \begin{Vmatrix} 0 & S_1 \\ R_2 & 0 \end{Vmatrix}$$

a toutes ses valeurs propres réelles de module inférieur à un (i.e. $|\lambda(B)| < 1$) et que H est définie comme suit :

$$H = \begin{Vmatrix} E_1 & 0 \\ -R_2 & E_2 \end{Vmatrix}^{-1} = \begin{Vmatrix} E_1 & 0 \\ R_2 & E_2 \end{Vmatrix}. \tag{2.72}$$

Ce choix de H correspond à la méthode de surrelaxation successive (2.38) avec $\tau = 1$ (voir point 2.4.3). Il est évident que chaque valeur propre de la matrice

$$HA = \begin{Vmatrix} E_1 & -S_1 \\ 0 & E_2 - R_2S_1 \end{Vmatrix}$$

ou bien est égale à un, ou bien constitue une v.p. de la matrice $E_2 - R_2S_1$ dont toutes les valeurs propres sont réelles et positives. Il suffit de remarquer, pour avoir le dernier résultat, que toute v.p. de la matrice R_2S_1 est encore une v.p. de

$$B^2 = \begin{Vmatrix} S_1R_2 & 0 \\ 0 & R_2S_1 \end{Vmatrix}$$

(il y a positivité des valeurs propres de B^2 parce qu'elles sont les carrés de celles de B) et qu'on a les inégalités

$$\lambda(R_2S_1) = \lambda(B^2) = \lambda^2(B) < 1.$$

Ainsi, la matrice HA a toutes ses valeurs propres réelles et positives, ce qui permet d'optimiser la méthode (2.70) moyennant les résul-

tats ci-dessus relatifs au choix des paramètres de Tchebycheff. Il y a lieu de noter que sous l'hypothèse supplémentaire de A symétrique ($R_2 = S_1^*$), la matrice HA possède un système complet de vecteurs propres.

Avant de passer au point suivant, on évalue la vitesse de convergence de (2.70) avec paramètres de Tchebycheff lorsque les matrices A et H se définissent par (2.71) et (2.72) moyennant $p = p(A) = \frac{\beta(A)}{\alpha(A)}$ avec $p \gg 1$. Les résultats du point 4.2.3 et la forme de HA aidant, on constate facilement que

$$\alpha(HA) = 1 - \beta(B^2) = 1 - \beta^2(B) = 1 - (1 - \alpha(A))^2 = \alpha(A)(2 - \alpha(A)) \approx 2\alpha(A),$$
$$B(HA) = 1,$$

donc

$$P(HA) = \frac{\beta(HA)}{\alpha(HA)} \approx \frac{1}{2\alpha(A)} \approx \frac{1}{4}\frac{\beta(A)}{\alpha(A)} = \frac{1}{4}p,$$

d'où pour (2.70), par suite de (2.69),

$$R(\tilde{T}_{\tau_{opt}}) \approx \frac{4}{\sqrt{p}}, \qquad W \approx W_0 |\ln \varepsilon| \frac{\sqrt{p}}{4}. \tag{2.73}$$

Ainsi, la vitesse de convergence asymptotique de la variante considérée est égale à celle de la surrelaxation successive avec τ_{opt}.

4.2.5. Comparaison de la vitesse de convergence des méthodes itératives pour les systèmes d'équations aux différences

Nous avons jusqu'à présent évalué la vitesse de convergence des méthodes itératives par recours au nombre de conditionnement de la matrice A. Appliquons ces méthodes aux systèmes d'équations aux différences finies qui approchent des équations elliptiques en dimension deux. Soit le système d'équations aux différences

$$\varphi_{k,l} = a_{k,l}\varphi_{k-1,l} + b_{k,l}\varphi_{k,l-1} + c_{k,l}\varphi_{k+1,l} + d_{k,l}\varphi_{k,l+1} + f_{k,l}, \tag{2.74}$$

où (k, l) parcourent un ensemble d'indices Q.

Supposons qu'il correspond à chaque nœud du réseau de discrétisation une valeur $\varphi_{k,l}$ et une équation, i.e. il y a autant d'équations et d'inconnues que de nœuds. Pour mettre le système sous forme matricielle on numérote les nœuds et on ordonne en conséquence les composantes de $\{\varphi_{k,l}\}$, puis les équations.

Prenons le système (2.74) basé sur un réseau carré, dont la solution est définie dans le carré unité. On a n, $l = 1, \ldots, m$ dans (2.74) et la matrice A du système

$$A\varphi = f \tag{2.75}$$

est d'ordre $n = m^2$.

Partageons toutes les composantes en deux groupes selon que $k + l$ sont paires ou impaires et numérotons séparément les éléments de chaque groupe. La fig. 14 représente le cas $m = 3$.

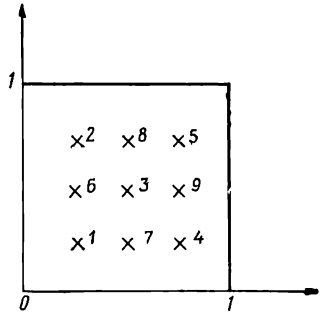


Fig. 14

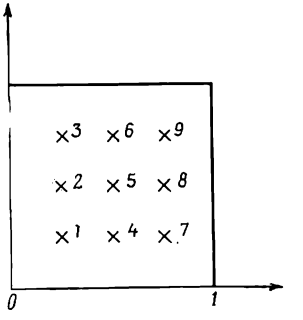


Fig. 15

La matrice correspondante $A = A_1$ de (2.75) est définie par la relation (qui est juste même pour m quelconque)

$$A_1 = \begin{Vmatrix} E_1 & -S_1 \\ -R_2 & E_2 \end{Vmatrix}, \tag{2.76}$$

l'ordre de E_1 étant égal au nombre de termes du premier groupe et l'ordre de E_2 à celui des termes du second.

Un autre procédé de numérotage pour $m = 3$ est visualisé par la fig. 15.

On partage les composantes en trois groupes selon la valeur de k. La matrice correspondante $A = A_2$ de (2.75) s'écrit pour $m = 3$

$$A_2 = \begin{Vmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{Vmatrix}. \tag{2.77}$$

On multiplie le système (2.75) à cette A par D^{-1}, où

$$D = \begin{Vmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A_{33} \end{Vmatrix} \tag{2.78}$$

est diagonale par blocs, et on obtient le système

$$A_3\varphi \equiv \begin{Vmatrix} E_1 & -S_1 & 0 \\ -R_2 & E_2 & -S_2 \\ 0 & -R_3 & E_3 \end{Vmatrix}, \qquad \varphi = F, \tag{2.79}$$

avec $F = D^{-1}f$, $S_l = -A_{ll}^{-1}A_{l,\,l+1}$, $R_l = -A_{ll}^{-1}A_{l,\,l-1}$, $l = 1, 2, 3$.

Avec $k = m$ quelconque, la matrice A_3 est de la forme (2.32).

Si les opérateurs des problèmes elliptiques de départ sont symétriques et définis positifs et si l'approximation s'effectue ou bien par des relations intégrales du bilan, ou bien par des méthodes variationnelles, alors les matrices construites A possèdent un système complet de vecteurs propres et leurs valeurs propres sont positives. Dans des cas exceptionnels (tels que le domaine rectangulaire et les coefficients constants), on calcule exactement les bornes du spectre (comme on l'a fait dans le Chapitre premier pour le laplacien discrétisé basé sur un carré) ou on les évalue (trop grossièrement des fois), ou encore on les cherche par une méthode itérative (Ljusternik, Lanczos, etc.).

Seul nous intéressera le cas où la matrice A du système (2.75) provient de l'approximation du problème de Dirichlet pour l'équation de Poisson dans le carré unité par des relations aux différences ordinaires à cinq points avec un réseau carré de pas $h \ll 1$. Ceci étant, nous nous bornerons à A de la forme (2.76).

Tableau 1

Nom de la méthode	Vitesse de convergence asymptotique $R(T)$
Méthode itérative simple	$\frac{2}{p} \approx \frac{\pi^2 h^2}{2}$
Méthode itérative cyclique de Tchebycheff	$\frac{2}{\sqrt{p}} \approx \pi h$
Surrelaxation successive	$\frac{4}{\sqrt{p}} \approx 2\pi h$
Méthode itérative cyclique de Tchebycheff (2.70)-(2.72)	$\frac{4}{\sqrt{p}} \approx 2\pi h$

On trouve par des calculs directs

$$\begin{aligned} \alpha(A) &= 2\sin^2\frac{\pi h}{2} \approx \frac{\pi^2 h^2}{2}, \\ \beta(A) &= 2 - \alpha(A) \approx 2 - \frac{\pi^2 h^2}{2}, \\ p &= p(A) = \frac{\beta(A)}{\alpha(A)} \approx \frac{4}{\pi^2 h^2}. \end{aligned} \tag{2.80}$$

Résumons certains résultats sous forme de tableau sans oublier que la matrice A du système (2.75) est mal conditionnée (car $h \ll 1$) (voir p. 173).

4.3. Méthodes itératives non stationnaires

On étudiera dans ce numéro des méthodes itératives non stationnaires qui consistent à minimiser de proche en proche une fonctionnelle quadratique.

4.3.1. Des théorèmes de convergence

Soient le système d'équations algébriques linéaires

$$A\varphi = f \tag{3.1}$$

et la fonctionnelle quadratique *)

$$J(\varphi) = (D(\varphi - \varphi^*), \varphi - \varphi^*), \tag{3.2}$$

où $\varphi^* = A^{-1}f$ est la solution exacte de (3.1) et D une matrice définie positive symétrique.

Comme $J(\varphi) > 0$ pour tout $\varphi \neq \varphi^*$ et $J(\varphi^*) = 0$, résoudre le système (3.1) équivaut à minimiser la fonctionnelle (3.2), i.e. à trouver le vecteur φ^* qui rend $J(\varphi)$ minimale.

Si

$$D = A^*A,$$

alors

$$J(\varphi) = (A\varphi - f, A\varphi - f) = \| A\varphi - f \|_2^2 \tag{3.3}$$

constitue la *fonctionnelle du carré du résidu*.

Si l'on fait l'hypothèse « A est une matrice définie positive symétrique et $D = A$ », alors

$$J(\varphi) = J_1(\varphi) + (A\varphi^*, \varphi^*),$$

où

$$J_1(\varphi) = (A\varphi, \varphi) - 2(\varphi, f). \tag{3.4}$$

Ainsi, $J(\varphi)$ est égale à la constante $(A\varphi^*, \varphi^*) = (f, \varphi^*)$ près à la fonctionnelle variationnelle connue $J_1(\varphi)$. On note l'indépendance de (3.3) et (3.4) vis-à-vis du vecteur cherché φ^*.

Ce que nous venons de dire permet d'optimiser les méthodes itératives en termes de Calcul variationnel. Soit la méthode itérative

$$\varphi^{j+1} = \varphi^j - H_\tau(A\varphi^j - f) \tag{3.5}$$

dont la matrice H_τ dépend des paramètres $\tau_1, \ldots, \tau_s$. On suppose sa convergence pour $\tau_1, \ldots, \tau_s$ d'un ensemble Q, la suite $\{\varphi^j\}$, réalisant la minimisation successive de la fonctionnelle (3.2). Dans

*) Le présent numéro s'inspire des résultats de Kouznétsov [XI; 6 a)].

le numéro précédent, les valeurs de $\tau_1, \ldots, \tau_s$ ont été choisies en minimisation du rayon spectral $\beta(T_\tau)$ de la matrice de transition $T_\tau = E - H_\tau A$, et, dans les méthodes qui suivent, le choix de ces paramètres est régi par la condition de minimum à chaque pas de la fonctionnelle (3.2).

La matrice D étant définie positive, on voit aisément que la relation

$$\| \psi \|_D = (D\psi, \psi)^{1/2} \tag{3.6}$$

définit une norme sur l'espace des vecteurs erreurs et qu'une matrice B a pour norme

$$\| B \|_D = \sup_{\psi \neq 0} \frac{\| B\psi \|_D}{\| \psi \|_D}. \tag{3.7}$$

(On dit parfois « D-norme ».)

Nous avons supposé qu'avec $(\tau_1, \ldots, \tau_s) \in Q$, la fonctionnelle (3.2) à une matrice D est minimisée de proche en proche, i.e.

$$\| T_\tau \psi^j \|_D < \| \psi^j \|_D \tag{3.8}$$

quel que soit $\psi^j = \varphi^j - \varphi^*$.

S'agissant d'une méthode stationnaire T_τ est constante, si bien que la dernière inégalité entraîne

$$\| T_\tau z^j \|_D < 1,$$

$z^j = \psi^j / \| \psi^j \|_D$, pour tout vecteur $\psi^j \neq 0$. Comme $v = \{z: \| z \|_D = 1\}$ est un ensemble fermé borné *),

$$\sup_{\| z \|_D = 1} \| T_\tau z \|_D = \| T_\tau \|_D$$

est réalisé par un vecteur z_0. Compte tenu de (3.8), on a

$$\| T_\tau \|_D = \| T_\tau z_0 \|_D < 1. \tag{3.9}$$

Ainsi,

$$\| T_\tau \|_D < 1$$

sous l'hypothèse $(\tau_1, \ldots, \tau_s) \in Q$.

L'analogue non stationnaire de (3.5) s'écrit

$$\varphi^{j+1} = \varphi^j - H_j (A\varphi^j - f), \tag{3.10}$$

où $H_j = H(\tau_1^{(j)}, \ldots, \tau_s^{(j)})$ et les paramètres $\{\tau_i^{(j)}\}$ vérifient l'équation

$$J(\varphi^j - H_j(A\varphi^j - f)) = \inf_{\tau_1, \ldots, \tau_s} J(\varphi^j - H_\tau (A\varphi^j - f)). \tag{3.11}$$

Je dis qu'avec les hypothèses faites sur la méthode (3.5) (i.e. l'ensemble Q n'est pas vide), la méthode itérative (3.10), (3.11) est

*) Les notations de la forme $v = \{z: \| z \|_D = 1\}$ signifient que l'ensemble v est formé d'éléments z tels que $\| z \|_D = 1$.

convergente et

$$\| \widetilde{T} \|_D \leqslant \| T_\tau \|_D \tag{3.12}$$

quel que soit $\tau = (\tau_1, \ldots, \tau_s) \in Q$, $\widetilde{T}$ étant l'opérateur de transition de (3.10), (3.11).

Il suffit de démontrer (3.12). Supposons, en effet, la méthode (3.5) convergente pour $\tau_1, \ldots, \tau_s$ et telle que $\| T_\tau \psi^j \|_D < \| \psi^j \|_D$ quel que soit ψ^j. Il est déjà connu que dans ce cas

$$\| T_\tau \|_D < 1.$$

Par définition,

$$\| T_j \psi^j \|_D \leqslant \| T_\tau \psi^j \|_D,$$

si bien que

$$\frac{\| T_j \psi^j \|_D}{\| \psi^j \|_D} \leqslant \| T_\tau \|_D < 1$$

pour tout ψ^j, donc

$$\| \widetilde{T} \| = \sup_{\psi^j \neq 0} \frac{\| T_j \psi^j \|_D}{\| \psi^j \|_D} \leqslant \| T_\tau \|_D < 1,$$

c.q.f.d.

On démontre une autre affirmation après avoir défini la vitesse de convergence asymptotique de la méthode non stationnaire par la relation *)

$$R(\widetilde{T}) = -\left[\lim_{j \to \infty} \frac{1}{j} \ln \left\{ \sup_{\xi^0 \neq 0} \frac{\| \xi^j \|_D}{\| \xi^0 \|_D} \right\}\right] = -\lim_{j \to \infty} \left[\frac{1}{j} \ln \| \widetilde{T}^j \|_D \right].$$

Si

$$\| T_\tau \|_D = \beta (T_\tau)$$

pour n'importe quels $\tau_1, \ldots, \tau_s \in Q$ tels qu'il y ait convergence de la méthode (3.5), alors la vitesse de convergence asymptotique du procédé itératif non stationnaire associé est au moins égale à celle de la méthode stationnaire avec paramètres optimaux.

En effet, si l'on est dans les conditions indiquées, alors

$$R(T_\tau) = -\ln \beta (T_\tau)$$

(voir point 4.2.2). D'autre part,

$$\left[\frac{\left\| \prod_{j=1}^{k} T_j \psi^0 \right\|_D}{\| \psi^0 \|_D} \right]^{1/k} \leqslant \| \widetilde{T} \|_D \leqslant \| T_\tau \|_D.$$

D'où

$$R(\widetilde{T}) \geqslant R(T_\tau). \tag{3.13}$$

Cette inégalité est vraie pour tout $\tau_1, \ldots, \tau_s \in Q$, d'où le résultat cherché.

*) Comme dans le cas stationnaire, $R(\widetilde{T})$ ne change pas avec la norme, i.e. avec le choix de la matrice D.

Il en découle qu'avec les paramètres choisis en minimisation de la D-norme de l'opérateur de transition T_τ, l'optimisation variationnelle garantit une plus grande vitesse de convergence.

4.3.2. Méthode du plus petit résidu

Prenons

$$D = A^*A \text{ et } H = \tau E \tag{3.14}$$

et soit $A = A^* > 0$.

Il est connu dès le point 4.2.1 que la méthode itérative

$$\varphi^{j+1} = \varphi^j - \tau (A\varphi^j - f) \tag{3.15}$$

est alors convergente pour $\tau \in \left(0, \frac{2}{\beta(A)}\right)$, et l'analyse spectrale élémentaire montre que

$$\| E - \tau A \|_D = \| E - \tau A \|_2 = \beta (E - \tau A). \tag{3.16}$$

Considérons le processus non stationnaire correspondant qui constitue la *méthode du plus petit résidu* :

$$\varphi^{j+1} = \varphi^j - \tau_j (A\varphi^j - f), \tag{3.17}$$

où

$$\tau_j = \frac{(A\xi^j, \xi^j)}{(A\xi^j, A\xi^j)} = \frac{(A\xi^j, \xi^j)}{\| A\xi^j \|_2^2} \tag{3.18}$$

($\xi^j = A\varphi^j - f$ est le vecteur résidu). Il résulte de (3.17)

$$\|(E - \tau_j A)\psi^j\|_D^2 = \|(E - \tau_j A)\xi^j\|_2^2 = \inf_\tau J(\varphi^j - \tau\xi^j) =$$

$$= \inf_\tau \| (E - \tau A)\xi^j \|_2^2 = \inf_\tau \{ \| \xi^j \|_2^2 - 2\tau (A\xi^j, \xi^j) + \tau^2 \| A\xi^j \|_2^2 \}.$$

On trouve évidemment τ_j à l'aide de

$$\frac{\partial}{\partial \tau} \{ \| \xi^j \|_2^2 - 2\tau (A\xi^j, \xi^j) + \tau^2 \| A\xi^j \|_2^2 \} = 0.$$

La méthode (3.15) converge pour $\tau \in \left(0, \frac{2}{\beta}\right)$, $\beta = \beta(A)$, si bien qu'il en est également, selon le n° 4.2, de la méthode proposée, et

$$\| \tilde{T} \| \leqslant \frac{\beta - \alpha}{\beta + \alpha}. \tag{3.19}$$

Il est connu (Faddeev, Faddeeva [VIII] ; Marchouk, Kouznétsov [VIII ; 12]) que (3.19) a lieu avec le signe égalité.

Ecrivons la relation pour la suite des normes des résidus de (3.17)-(3.18) :

$$\| \xi^{j+1} \|_2^2 = \| \xi^j \|_2^2 - \frac{(A\xi^j, \xi^j)^2}{\| A\xi^j \|_2^2}.$$

On voit que pour que cette suite soit monotone décroissante dans l'espace des vecteurs réels, il suffit que la matrice A soit définie positive dans cet espace (la symétrie peut être négligée). Cette propriété découverte par Kouznétsov [XI; 6] a inspiré un théorème de convergence pour les méthodes itératives non stationnaires et, en particulier, la démonstration de la convergence de la méthode étudiée pour les systèmes d'équations à matrices définies positives non symétriques. La question est examinée en détail dans la monographie de Marchouk et Kouznétsov [VIII; 12].

La méthode du plus petit résidu remonte à l'article de Krasnosselski et Krein [XI; 5].

Elle possède la propriété, importante dans la pratique, de converger aux premières itérations beaucoup plus vite qu'à la vitesse asymptotique. Forsythe et Motzkin [XI] ont noté la même circonstance pour la technique de la plus grande pente (voir Kantorovitch [XI; 4]). Ils ont montré que dans ce cas l'erreur de méthode est asymptotiquement une combinaison linéaire de deux vecteurs propres de A associés aux valeurs propres $\alpha(A)$ et $\beta(A)$. Il en est de même de la méthode du plus petit résidu. Ainsi, les méthodes itératives non stationnaires jouissent, à l'instar des procédés stationnaires, de la propriété asymptotique de perdre progressivement de la vitesse de convergence (pour les paramètres fixes).

On accélère la convergence de la méthode du plus petit résidu par recours périodique à un pas de la variante à trois niveaux procédant selon les formules

$$\varphi^{j+1} = \varphi^j - \tau_j (A\varphi^j - f) - \gamma_j A (A\varphi^j - f), \tag{3.20}$$

avec τ_j et γ_j pris pour solution du système de deux équations

$$\begin{cases} \dfrac{\partial}{\partial \tau_j} \| \xi^j - \tau_j A\xi^j - \gamma_j A^2 \xi^j \|_2^2 = 0, \\ \dfrac{\partial}{\partial \gamma_j} \| \xi^j - \tau_j A\xi^j - \gamma_j A^2 \xi^j \|_2^2 = 0. \end{cases} \tag{3.21}$$

4.3.3. Méthode du gradient conjugué

On définit dans l'espace primitif des vecteurs une D-norme et un sous-espace G_s de base $\{g_i\}_{i=1}^s$. Le problème d'approcher au mieux la solution $\varphi^* = A^{-1}f$ du système $A\varphi = f$ sur la variété $U_s^0 = \varphi^0 + G_s = \{\varphi: \varphi = \varphi^0 + \psi, \psi \in G_s\}$ consiste alors à chercher un vecteur $\hat{\psi} \in G_s$ tel qu'on ait

$$\| \varphi^* - (\varphi^0 + \hat{\psi}) \|_D = \min_{\psi \in G_s} \| \varphi^* - (\varphi^0 + \psi) \|_D =$$

$$= \min_{\alpha_1, \alpha_2, \ldots, \alpha_s} \Big\| \varphi^* - \varphi^0 - \sum_{i=1}^{s} \alpha_i g_i \Big\|_D. \tag{3.22}$$

Le système d'équations pour déterminer les coefficients $\{\alpha_i^*\}$ du développement $\hat{\psi} = \sum_{i=1}^{s} \alpha_i^* g_i$ s'écrit

$$B\alpha = F,$$

avec $B = (b_{ij})$ une matrice d'ordre s d'éléments

$$b_{ij} = (g_i, g_j)_D, \quad i, j = 1, \ldots, s,$$

et $F = (F_1, \ldots, F_s)$ un vecteur composé de

$$F_i = (\varphi^* - \varphi^0, g_i)_D, \quad i = 1, \ldots, s.$$

Il découle des résultats précédents que le cas

$$(Dg_i, g_j) = \delta_{ij} \| g_i \|_D^2$$

(δ_{ij} étant le symbole de Kronecker) est le plus simple (du point de vue de la mise en œuvre), i.e. $\{g_i\}$ est une base orthogonale par rapport à D de l'espace G_s. Si l'on est dans la dernière condition, alors

$$\alpha_i^* = \frac{(\varphi^* - \varphi^0, g_i)_D}{\| g_i \|_D^2} = \frac{(D(\varphi^* - \varphi^0), g_i)}{\| g_i \|_D^2}, \quad i = 1, \ldots, s. \tag{3.23}$$

Une condition suffisante de (3.23) est de connaître le vecteur $D\varphi^*$. On la remplit par exemple ou bien quand $D = A$ avec $A = A^* > 0$, ou bien lorsque $D = A^*A$ avec A quelconque.

Particularisons le problème variationnel (3.22) afin d'utiliser, donc d'étudier, une classe de méthodes, à savoir on suppose que G_s constitue l'enveloppe linéaire du système libre $\{A^i(\varphi^0 - \varphi^*)\}_{i=1}^s = \{A^{i-1}(A\varphi^0 - f)\}_{i=1}^s$ et que la matrice A est symétrique et définie positive. On sait que si le sous-espace donné G_s possède une base $\{g_i\}_{i=1}^s$ orthogonale par rapport à A, alors on cherche l'approximation $\hat{\varphi}$ de φ^* par les formules

$$\hat{\varphi} = \varphi^0 + \sum_{i=1}^{s} \alpha_i g_i,$$

$$\alpha_i = \frac{(\varphi^* - \varphi^0, g_i)_A}{(g_i, g_i)_A} = -\frac{(A\varphi^0 - f, g_i)}{(g_i, Ag_i)}, \quad i = 1, \ldots, n, \tag{3.24}$$

ce qui s'écrit également

$$\varphi^k = \varphi^{k-1} - \alpha_k g_k,$$

$$\alpha_k = \frac{(\xi^{k-1}, g_k)}{(Ag_k, g_k)}, \tag{3.25}$$

$$k = 1, \ldots, s,$$

avec $\xi^k = (A\varphi^k - f)$ le vecteur résidu et $\hat{\varphi} = \varphi^s$.

Le procédé de Schmidt est le plus utilisé pour construire une base dans des sous-espaces du type G_s. Mais il exige, pour A d'ordre élevé,

un grand nombre d'opérations arithmétiques et sa mise en œuvre numérique occupe beaucoup de mémoires. Quand les matrices sont symétriques sans être définies positives, on construit efficacement une base A^2-orthogonale ($D = A^2$) de G_s par la méthode des itérations minimisées de Lanczos pour la description de laquelle nous renvoyons à la monographie de Faddeev et Faddeeva [VIII]. Le procédé le plus économique pour orthogonaliser les vecteurs $\{A^{i-1}(A\varphi^0 - f)\}_{i=1}^s$ par rapport à A dans le cas des matrices symétriques et définies positives est la méthode du gradient conjugué régie par les formules

$$g_k = \begin{cases} \xi^0 & \text{si } k=1, \\ \xi^{k-1} - b_k g_{k-1} & \text{si } k>1, \end{cases}$$
$$b_k = \frac{(A\xi^{k-1}, g_{k-1})}{(Ag_{k-1}, g_{k-1})}, \qquad \varphi^k = \varphi^{k-1} - \alpha_k g_k, \tag{3.26}$$
$$\alpha_k = \frac{(\xi^{k-1}, g_k)}{(Ag_k, g_k)}, \qquad k=1, \ldots, s,$$

$\{\xi^k = A\varphi^k - f\}_{k=1}^s$ étant les vecteurs résidus.

On démontre que les vecteurs $\{g_k\}_{k=1}^s$ ainsi formés constituent une base orthogonale par rapport à A de l'espace G_s si $\{A^{k-1}\xi^0\}_{k=1}^s$ sont linéairement indépendants. On montre d'abord la non-nullité des $\{g_k\}_{k=1}^s$. En effet, si l'on effectue une élimination successive dans (3.26), on voit que

$$g_k = A^{k-1}\xi^0 + \sum_{i=1}^{k-1} \beta_{ki} A^{i-1}\xi^0, \qquad k=1, \ldots, s, \tag{3.27}$$

avec certains $\{\beta_{ki}\}$, si bien que $g_k = 0$ contredit la condition d'indépendance linéaire des vecteurs $\{A^{k-1}\xi^0\}_{k=1}^s$.

Passons à la propriété de A-orthogonalité des $\{g_k\}_{k=1}^s$ et supposons la validité des relations

$$\begin{aligned} (Ag_k, g_j) &= 0, \quad j = 1, \ldots, k-1, \\ (\xi^k, g_j) &= 0, \quad j = 1, \ldots, k, \\ \alpha_j &> 0, \quad j = 1, \ldots, k, \end{aligned} \tag{3.28}$$

pour un $k \geqslant 2$ (lorsque $k = 1, 2$, on les vérifie directement). Démontrons-les au $(k+1)$-ième pas. Nous aurons besoin des égalités

$$Ag_j = \frac{1}{\alpha_j}[g_j + b_j g_{j-1} - g_{j+1} - b_{j+1} g_j] =$$
$$= \varepsilon_j g_{j+1} + \beta_j g_j + \gamma_j g_{j-1}, \qquad j=1, \ldots, k-1 \tag{3.29}$$

(on suppose $g_0 = 0$) qu'on établit moyennant les égalités obtenues à partir de (3.26) et les conditions

$$\xi^j = \xi^{j-1} - \alpha_j A g_j,$$
$$g_{j+1} = \xi^j - b_{j+1} g_j, \qquad j=1, \ldots, k-1.$$

Comme

$$(Ag_{k+1}, g_j) = (A\xi^k - b_{k+1}Ag_k, g_j) = (A\xi^k, g_j) - b_{k+1}(Ag_k, g_j) = \\ = (\xi^k, Ag_j) = (\xi^k, \varepsilon_j g_{j+1} + \beta_j g_j + \gamma_j g_{j-1}) = 0, \\ j = 1, \ldots, k-1,$$

par hypothèse et $(Ag_{k+1}, g_k) = 0$ par construction, on a

$$(Ag_{k+1}, g_j) = 0 \tag{3.30}$$

quels que soient $j \leqslant k$. Du moment que $(\xi^{k+1}, g_{k+1}) = 0$, il résulte par construction de (3.28)

$$(\xi^{k+1}, g_j) = (\xi^k, g_j) - \alpha_{k+1}(Ag_{k+1}, g_j), \qquad j = 1, \ldots, k,$$

si bien que

$$(\xi^{k+1}, g_j) = 0 \tag{3.31}$$

pour tout $1 \leqslant j \leqslant k+1$. Réunissons (3.30) et (3.31) et prenons en considération

$$\alpha_{k+1} = \frac{(\xi^k, g_{k+1})}{(Ag_{k+1}, g_{k+1})} = \frac{(\xi^k, \xi^k)}{(Ag_{k+1}, g_{k+1})} > 0,$$

on aboutit à la validité au $(k+1)$-ième pas de toutes les relations (3.28). En poursuivant le processus de récurrence jusqu'au pas s, on établit l'orthogonalité par rapport à A du système $\{g_k\}$. Ainsi, le problème variationnel (3.22) se trouve résolu par la méthode du gradient conjugué.

Voyons maintenant un cas de dégénérescence où le système $\{A^{i-1}\xi^0\}_{i=1}^k$ est libre pour un $k \geqslant 1$ et le système $\{A^{j-1}\xi^0\}_{i=1}^{k+1}$ ne l'est plus, i.e.

$$-\sum_{j=0}^{k} C_j A^j \xi^0 \equiv \sum_{j=0}^{k} C_j A^{j+1}(\varphi^* - \varphi^0) = 0 \tag{3.32}$$

avec certains $\{C_j\}_{j=0}^k$ non tous nuls. Le coefficient $C_0 \neq 0$ sinon la multiplication de (3.32) par A^{-1} donne

$$\sum_{j=0}^{k} C_j A^j (\varphi^* - \varphi^0) = 0,$$

ce qui contredit l'indépendance linéaire de $\{A^j(\varphi^* - \varphi^0)\}_{j=1}^k$. Il découle de (3.32)

$$\varphi^* - \varphi^0 = -\frac{1}{C_0} A^{-1} \sum_{j=1}^{k} C_j A^{j+1}(\varphi^* - \varphi^0) = \sum_{j=1}^{k} \frac{C_j}{C_0} A^{j-1}\xi^0,$$

autrement dit, $\varphi^* - \varphi^0 \in G_s$. La dernière relation et (3.28) entraînent que $\hat{\psi}$ est nécessairement égal au vecteur $\varphi^* - \varphi^0$ ($\varphi^* = \varphi^0 +$

n

$+ \hat{\psi} = \varphi^0 + \varphi^k$), i.e. le procédé utilisé fournit la solution exacte du système (3.1) dès le k-ième pas.

La méthode du gradient conjugué étant un procédé d'orthogonalisation permet d'accélérer la convergence des méthodes itératives stationnaires. S'agissant par exemple du processus itératif

$$\varphi^{k+1} = \varphi^k - B(A\varphi^k - f), \qquad k = 1, 2, \ldots, \tag{3.33}$$

à matrices A et B symétriques et définies positives, les formules d'accélération correspondantes s'écrivent

$$\begin{aligned} g_k &= \begin{cases} B\xi^0 & \text{si } k=1, \\ B\xi^{k-1} - b_k g_{k-1} & \text{si } k>1, \end{cases} \\ b_k &= \frac{(AB\xi^{k-1}, g_{k-1})}{(Ag_{k-1}, g_{k-1})}, \\ \varphi^k &= \varphi^{k-1} - \alpha_k g_k, \\ \alpha^k &= \frac{(\xi^{k-1}, g_k)}{(Ag_k, g_k)}, \\ & k=1, 2, \ldots, s. \end{aligned} \tag{3.34}$$

La technique du gradient conjugué est, en théorie, une méthode directe parce qu'avec $s > n$, n étant l'ordre de A, le système $\{A^i\xi^0\}_{i=0}^{s-1}$ est toujours libre, si bien que le processus aboutit pour un $k \leqslant s$ à la solution exacte. D'autre part, sa mise en œuvre sur ordinateur présente pour des matrices d'ordre élevé l'inconvénient de donner lieu, après plusieurs dizaines d'itérations, à l'instabilité numérique de l'orthogonalisation et le processus réel ne reflète plus les propriétés de la méthode utilisée. On assimile donc la méthode du gradient conjugué à un procédé itératif non stationnaire s-cyclique (i.e. tous les s pas on initialise avec une nouvelle approximation de départ). On évalue maintenant la vitesse de convergence par l'intermédiaire de celle de la méthode cyclique de Tchebycheff. En effet, on a pour $s \geqslant 1$

$$\min_{\alpha_1, \ldots, \alpha_s} \frac{\left\| \varphi^* - \varphi^0 - \sum_{i=1}^{s} \alpha_i A^{i-1}\xi^0 \right\|_A^2}{\| \varphi^* - \varphi^0 \|_A^2} \leqslant \| \tilde{T}_s \|_A \leqslant$$

$$\leqslant \left\| E - \sum_{i=1}^{s} \beta_i A^i \right\|_2^2 = \left[\beta \left(E - \sum_{i=1}^{s} \beta_i A^i \right) \right]^2, \tag{3.35}$$

où $\tilde{T}_s$ est l'opérateur qui transforme le vecteur erreur sur s pas de la méthode du gradient conjugué et $\{\beta_i\}_{i=1}^s$ sont des nombres réels

arbitraires. Posons en particulier

$$E - \sum_{i=1}^{s} \beta_i A^i = \prod_{i=1}^{s} (E - \tau_i A) \tag{3.36}$$

et utilisons la majoration (3.12) ci-dessus, il vient que le gradient conjugué s-cyclique converge au moins comme la méthode de Tchebycheff avec le même s et qu'on a donc dans les deux cas les mêmes estimations de la vitesse de convergence. Il y a lieu de parler de deux particularités numériques de la méthode du gradient conjugué. *Primo*, elle demande à chaque pas plus (beaucoup plus des fois) d'opérations arithmétiques et logiques que la méthode itérative de Tchebycheff (on le prendra en considération lorsqu'on aura à opter pour telle ou telle méthode). *Secundo*, sa convergence en norme correspondante est sensiblement plus vite (surtout aux premières itérations) que l'estimation ne le laisse voir. Ecrivons par exemple la relation qui découle de (3.22) pour la méthode du gradient conjugué (le procédé de définition de G_s) *) :

$$(A\psi^k, \psi^k) \leqslant \max_{m \leqslant \lambda \leqslant M} | \lambda P_k^2(\lambda) | \cdot (\psi^0, \psi^0), \tag{3.37}$$

où ψ^k est le vecteur erreur commise au k-ième pas, ψ^0 le vecteur erreur initiale et $P_k(\lambda)$ un polynôme de degré k en λ tel que $P_k(0) = 1$. Si l'on pose $M = 1$ (la chose est possible à condition de normaliser A de façon convenable) et $m = 0$ (sans restreindre la généralité) et si l'on suppose

$$P_k(\lambda) = (-1)^k \frac{\cos[(2k+1) \arccos \sqrt{\lambda}]}{(2k+1)\sqrt{\lambda}}, \tag{3.38}$$

on obtient la majoration

$$(A\psi^k, \psi^k) \leqslant \frac{(\psi^0, \psi^0)}{(2k+1)^2}. \tag{3.39}$$

L'égalité $m = 0$ fait que (3.39) reste en vigueur pour A dégénérée.

4.4. Méthode de décomposition

La méthode itérative des directions alternées proposée par Douglas, Peaceman et Rachford trouve de nombreuses applications dans les problèmes stationnaires de la physique mathématique. On en connaît maints schémas de mise en œuvre et de nombreuses variantes. La méthode repose sur des processus de relaxation spéciaux qui permettent de ramener un gros problème à une suite de problèmes simples. Nous appellerons *méthode de décomposition* toute technique procédant de cette réduction, et c'est sous cet angle que nous allons examiner la méthode classique des directions alternées.

*) Voir Iline [VIII; 6].

Prenons le système d'équations algébriques linéaires

$$A\varphi = f \tag{4.1}$$

avec la condition

$$A = A_1 + A_2, \tag{4.2}$$

A_1 et A_2 étant des matrices définies positives. Il est connu (voir n° 4.2) qu'on accélère sensiblement la convergence de la méthode sans augmenter en fait le nombre d'opérations par itération si l'on utilise au lieu de

$$\varphi^{j+1} = \varphi^j - \tau\,(A\varphi^j - f), \tag{4.3}$$

les itérations de la surrelaxation successive

$$B\,(\varphi^{j+1} - \varphi^j) = -(A\varphi^j - f) \tag{4.4}$$

à B dépendant du paramètre τ. L'inconvénient en est une contrainte sévère sur la forme de A et son spectre.

Dans une autre classe de méthodes, la matrice $B = B_\tau$ est définie par la relation

$$B_\tau = \frac{1}{2\tau}\,(E + \tau A) \tag{4.5}$$

et A est supposée symétrique et définie positive, auquel cas on a pour les vecteurs erreurs $\psi^j = \varphi^j - \varphi^*$

$$(E + \tau A)\,\psi^{j+1} = (E - \tau A)\,\psi^j$$

ou, ce qui revient au même,

$$\psi^{j+1} = T_\tau \psi^j,$$

où

$$T_\tau = (E + \tau A)^{-1}\,(E - \tau A) = (E - \tau A)\,(E + \tau A)^{-1} \tag{4.6}$$

est l'opérateur de transition de (4.4). Etant données la symétrie et la définie positivité de la matrice, on vérifie sans peine que T_τ est défini pour tout $\tau > 0$, symétrique et

$$\|T_\tau\|_2 = \beta\,(T_\tau) \max_{\lambda n(A)} \left|\frac{1 - \tau\lambda_n\,(A)}{1 + \tau\lambda_n\,(A)}\right|. \tag{4.7}$$

Ici

$$0 < \alpha = \alpha\,(A) \leqslant \lambda_n\,(A) \leqslant \beta = \beta\,(A). \tag{4.8}$$

Introduisons et étudions la fonction

$$q(\tau) = \max_{\lambda n(A)} |\,g\,(\tau, \lambda_n\,(A))\,|, \tag{4.9}$$

où

$$g\,(\tau, \lambda) = \frac{1 - \tau\lambda}{1 + \tau\lambda}. \tag{4.10}$$

Il est évident que $q(\tau) \geqslant 1$ pour $\tau \leqslant 0$, d'où la divergence de la méthode (4.4). Une condition nécessaire de convergence est donc $\tau > 0$. Supposons que $\tau > 0$. Comme

$$g'_\lambda(\tau, \lambda) = -\frac{2\tau}{(1+\tau\lambda)^2} < 0$$

pour τ, $\lambda > 0$ et $\lambda_n(A) \in [\alpha, \beta]$, on a

$$q(\tau) = \max_{\lambda_n(A)} |g(\tau, \lambda)| = \max\{|g(\tau, \alpha)|, |g(\tau, \beta)|\} =$$

$$= \max\left\{\left|\frac{1-\tau\alpha}{1+\tau\alpha}\right|, \left|\frac{1-\tau\beta}{1+\tau\beta}\right|\right\} < 1. \qquad (4.11)$$

D'où la convergence de (4.4) quel que soit $\tau > 0$.

La méthode (4.4) est une technique itérative stationnaire, si bien que l'optimisation consiste, selon le point 4.2.2, à minimiser $\beta(T_\tau)$ par rapport à τ, i.e. à résoudre le problème d'extrémum

$$q(\tau_{opt}) = \min_{\tau>0} q(\tau). \qquad (4.12)$$

Vu que

$$[g(\tau, \lambda)]'_\tau = -\frac{2\lambda}{(1+\tau\lambda^2)} < 0 \qquad (4.13)$$

quels que soient τ, $\lambda > 0$, on montre (comme au point 4.2.1 pour la méthode itérative simple) que la valeur $\tau = \tau_{opt}$ est la solution de l'équation

$$\frac{1-\tau\alpha}{1+\tau\alpha} = -\frac{1-\tau\beta}{1+\tau\beta} \qquad (4.14)$$

et est obtenue par la formule

$$\tau_{opt} = \frac{1}{\sqrt{\alpha\beta}}. \qquad (4.15)$$

En effet, si $\tau = \tau_{opt}$, alors

$$\frac{1-\tau_{opt}\alpha}{1+\tau_{opt}\alpha} = -\frac{1-\tau_{opt}\beta}{1+\tau_{opt}\beta} = \frac{1-\sqrt{\frac{\alpha}{\beta}}}{1+\sqrt{\frac{\alpha}{\beta}}} > 0,$$

donc

$$q(\tau_{opt}) = \frac{1-\sqrt{\frac{\alpha}{\beta}}}{1+\sqrt{\frac{\alpha}{\beta}}}. \qquad (4.16)$$

Si $\tau < \tau_{opt}$, alors on a, conformément à (4.13),

$$q(\tau) = \frac{1-\tau\alpha}{1+\tau\alpha} > q(\tau_{opt}).$$

On a enfin

$$q(\tau) = -\frac{1-\tau\beta}{1+\tau\beta} > q(\tau_{opt})$$

si $\tau > \tau_{opt}$.

On évalue la vitesse de convergence asymptotique de la méthode (4.4)-(4.5)

$$R(T_{\tau_{opt}}) = -\ln q(\tau_{opt}) \tag{4.17}$$

pour les matrices mal conditionnées, i.e. pour $p = p(A) \gg 1$. Selon (4.16),

$$\begin{aligned} q(\tau_{opt}) &\approx 1 - \frac{2}{\sqrt{p}}, \\ R(T_{\tau_{opt}}) &\approx \frac{2}{\sqrt{p}}. \end{aligned} \tag{4.18}$$

Les relations obtenues permettent de dire que la vitesse de convergence asymptotique de la méthode étudiée n'est que deux fois moins bonne que celle de la surrelaxation successive optimale. Cette méthode n'est pas sans mérites, mais on lui connaît également un défaut: la résolution du système

$$Bz^{j+1} = -\xi^j$$

($z^{j+1} = \varphi^{j+1} - \varphi^j$) peut exiger à chaque pas autant d'opérations arithmétiques que celle du système de départ. On a donc à remplacer $B = B_\tau$ par un opérateur « voisin » de façon qu'on conserve la vitesse de convergence et que le nombre d'opérations par itération soit comparable à celui de la surrelaxation successive. Dans la plupart des cas, on résout ce problème en choisissant

$$B = \frac{1}{2\tau}(E + \tau A_1)(E + \tau A_2), \tag{4.19}$$

et c'est ce B qu'on utilisera dans la suite.

Les schémas de décomposition font l'objet des ouvrages de Douglas, Peaceman, Rachford (voir [XV]), Samarski [III; 12] [XV; 11], Marchuk [XV], Yanenko [XV; 13], Diakonov [XV; 5] et de nombreux autres mathématiciens.

4.4.1. Cas commutatif

Supposons que les matrices A_1, A_2 de (4.2) sont symétriques et possèdent un système orthonormé commun de vecteurs propres $\{u_n\}$, i.e.

$$\begin{aligned} A_1 u_n &= \lambda_n(A_1)\, u_n, \\ A_2 u_n &= \lambda_n(A_2)\, u_n \end{aligned} \tag{4.20}$$

quel que soit n et le système $\{u_n\}$ forme une base orthonormée de l'espace initial des vecteurs. Développons un vecteur φ arbitraire par rapport au système $\{u_n\}$:

$$\varphi = \sum_n \varphi_n u_n, \qquad \varphi_n = (\varphi, u_n).$$

On constate sans peine que

$$A_1 A_2 \varphi = A_1 \left(\sum_n \lambda_n (A_2)\, \varphi_n u_n\right) =$$
$$= \sum_n \lambda_n (A_1)\, \lambda_n (A_2)\, \varphi_n u_n = A_2 \sum_n \lambda_n (A_1)\, \varphi_n u_n = A_2 A_1 \varphi.$$

L'arbitraire laissé sur φ permet d'en tirer la propriété de commutativité de A_1, A_2: $A_1 A_2 = A_2 A_1$.

Développons les vecteurs erreurs $\psi^j = \varphi^j - \varphi^*$ de la méthode de décomposition

$$(E + \tau A_1)\,(E + \tau A_2)\,(\varphi^{j+1} - \varphi^j) = -2\tau\,(A\varphi^j - f) \quad (4.21)$$

suivant le même système, il vient

$$\psi^j = \sum_n \psi_n^j u_n.$$

On obtient à l'aide de (4.21)

$$\psi_n^{j+1} = \frac{1-\tau\lambda_{1,\,n}}{1+\tau\lambda_{1,\,n}} \frac{1-\tau\lambda_{2,\,n}}{1+\tau\lambda_{2,\,n}} \psi_n^j, \quad (4.22)$$

avec la notation $\lambda_{i,\,n} = \lambda_n (A_i)$, $i = 1, 2$.

Il résulte de (4.22) (et, en général, de la commutativité des matrices A_1, A_2) que l'opérateur de transition de la méthode (4.21)

$$T_\tau = E - 2\tau\,(E + \tau A_2)^{-1}\,(E + \tau A_1)^{-1} A =$$
$$= (E + \tau A_2)^{-1}\,(E + \tau A_1)^{-1}\,(E - \tau A_1)\,(E - \tau A_2) \quad (4.23)$$

possède un système orthonormé complet de vecteurs propres $\{u_n\}$ et a ses valeurs propres réelles définies par

$$\lambda_n (T_\tau) = \frac{1-\tau\lambda_{1,\,n}}{1+\tau\lambda_{1,\,n}} \frac{1-\tau\lambda_{2,\,n}}{1+\tau\lambda_{2,\,n}}$$

moyennant les valeurs propres de A_1 et A_2. D'où la symétrie de la matrice T_τ, donc

$$\| T_\tau \|_2 = \beta (T_\tau) = \max_n \left| \frac{1-\tau\lambda_{1,\,n}}{1+\tau\lambda_{1,\,n}} \frac{1-\tau\lambda_{2,\,n}}{1+\tau\lambda_{2,\,n}} \right|. \quad (4.24)$$

L'égalité (4.24) signifie en termes de la théorie de l'optimisation des méthodes itératives stationnaires (voir point 4.2.2) que le choix du paramètre $\tau = \tau_{\text{opt}}$ à partir de la condition de minimum de

$$\beta (T_{\tau_{\text{opt}}}) = \min_\tau \beta (T_\tau) \quad (4.25)$$

est optimum non seulement sur le plan asymptotique, mais aussi pour tout $\varepsilon > 0$ si l'on demande de diminuer de $1/\varepsilon$ fois la norme euclidienne du vecteur erreur. On note de plus que, $\lambda_{1,\,n}$ et $\lambda_{2,\,n}$ étant positives, l'inégalité $\beta\,(T_\tau) < 1$ n'a lieu que pour $\tau > 0$, i.e. la positivité de τ constitue une condition nécessaire de convergence de la méthode de décomposition considérée.

Proposons-nous de résoudre le problème d'optimisation (4.25). Il s'agit en général d'un problème délicat qui reste ouvert, si bien qu'on le remplace par la recherche du minimum de la fonction

$$q^2(\tau) = \max_n \left| \frac{1-\tau\lambda_{1,\,n}}{1+\tau\lambda_{1,\,n}} \right| \max_n \left| \frac{1-\tau\lambda_{2,\,n}}{1+\tau\lambda_{2,\,n}} \right| \tag{4.26}$$

qui majore $\beta\,(T_\tau)$, l'approche suffisamment bien, voire lui est identique dans des cas particuliers.

On suppose que

$$\widetilde{\alpha} = \alpha\,(A_1) = \alpha\,(A_2), \quad \widetilde{\beta} = \beta\,(A_1) = \beta\,(A_2), \tag{4.27}$$

auquel cas

$$q^2(\tau) = [\max_{\widetilde{\alpha} \leqslant \lambda \leqslant \widetilde{\beta}} | g(\tau, \lambda) |]^2$$

car le plus grand module de

$$g(\tau, \lambda) = \frac{1-\tau\lambda}{1+\tau\lambda}$$

dépend seulement des bornes de l'intervalle $[\widetilde{\alpha}, \widetilde{\beta}]$ contenant $\lambda_{1,\,n}$ et $\lambda_{2,\,n}$ (on le sait dès le début du numéro). Le problème d'optimisation se ramène donc à la recherche de l'extrémum

$$q(\tau_{\text{opt}}) = \min_{\tau > 0} q(\tau), \tag{4.28}$$

problème qui a été déjà étudié. La solution est donnée par la formule (4.15) :

$$\tau_{\text{opt}} = \frac{1}{\sqrt{\widetilde{\alpha}\widetilde{\beta}}}, \tag{4.29}$$

d'où

$$\beta\,(T_{\tau_{\text{opt}}}) \leqslant q^2(\tau_{\text{opt}}) = \left[\frac{1-\sqrt{\frac{\widetilde{\alpha}}{\widetilde{\beta}}}}{1+\sqrt{\frac{\widetilde{\alpha}}{\widetilde{\beta}}}} \right]^2. \tag{4.30}$$

Puisque

$$\begin{aligned} \alpha &= \alpha\,(A) \geqslant \alpha\,(A_1) + \alpha\,(A_2) = 2\widetilde{\alpha}, \\ \beta &= \beta\,(A) \leqslant \beta\,(A_1) + \beta\,(A_2) = 2\widetilde{\beta}, \end{aligned} \tag{4.31}$$

donc

$$\frac{\widetilde{\alpha}}{\widetilde{\beta}} \leqslant \frac{\alpha}{\beta} < 1,$$

et la fonction

$$\varkappa(t) = \frac{1-\sqrt{t}}{1+\sqrt{t}}$$

décroît de façon monotone sur [0, 1] de 1 à 0, on a, compte tenu de (4.30),

$$\beta(T_{\tau_{\text{opt}}}) \leqslant \varkappa^2\left(\frac{\widetilde{\alpha}}{\widetilde{\beta}}\right) \leqslant \varkappa^2\left(\frac{\alpha}{\beta}\right) = \left[\frac{p-1}{p+1}\right]^2,$$

$p = \dfrac{\beta}{\alpha}$ étant le nombre de conditionnement de la matrice A.

On prend $p \gg 1$ et on obtient finalement les estimations

$$\begin{aligned} \beta(T_{\tau_{\text{opt}}}) &\leqslant 1 - \frac{4}{\sqrt{p}}, \\ R(T_{\tau_{\text{opt}}}) &\geqslant \frac{4}{\sqrt{p}}. \end{aligned} \tag{4.32}$$

Si l'on se rappelle les données du Tableau du point 4.2.5, on conclut que la vitesse de convergence est au moins tout aussi bonne que celle de la surrelaxation successive, la meilleure technique itérative stationnaire considérée. D'autre part, le procédé de décomposition proposé avec paramètre optimal converge au moins deux fois plus vite que (4.4)-(4.5) et est de mise en œuvre notablement plus simple car le processus (4.2) exige entre autres que les matrices $E + \tau A_i$, $i = 1, 2$, soient facilement inversibles.

L'un des points délicats de la méthode de décomposition est l'optimisation de schémas à plusieurs paramètres

$$(E + \tau_j A_1)(E + \tau_j A_2)(\varphi^{j+1} - \varphi^j) = -2\tau_j (A\varphi^j - f). \tag{4.33}$$

Le problème d'optimisation s'énonce comme suit. Etant donné $s \geqslant 1$, on demande la suite de paramètres $\{\tau_i\}_{i=1}^s$ qui minimisent une norme ou le rayon spectral de l'opérateur

$$T^{(s)} = \prod_{i=1}^{s} T_i, \tag{4.34}$$

où

$$T_i = (E + \tau_i A_2)^{-1}(E + \tau_i A_1)^{-1}(E - \tau_i A_1)(E - \tau_i A_2). \tag{4.35}$$

S'agissant des matrices commutatives, cela revient à chercher

$$\max_n \prod_{i=1}^{s} \left| \frac{1-\tau_i \lambda_{1,n}}{1+\tau_i \lambda_{1,n}} \frac{1-\tau_i \lambda_{2,n}}{1+\tau_i \lambda_{2,n}} \right| = \min_{\tau_1, \ldots, \tau_s}. \tag{4.36}$$

Comme dans le cas uniparamétrique, on ramène le problème à la recherche du minimum de la fonction

$$q(\tau_1, \ldots, \tau_s) = \max_{\alpha \leqslant \lambda \leqslant \beta} \prod_{i=1}^{s} \left| \frac{1-\tau_i\lambda}{1+\tau_i\lambda} \right|, \tag{4.37}$$

où

$$\alpha = \min(\alpha(A_1), \alpha(A_2)),$$

$$\beta = \max(\beta(A_1), \beta(A_2)).$$

Evidemment

$$\beta(T^{(s)}) \leqslant q^2(\tau_1, \ldots, \tau_s) < 1 \tag{4.38}$$

quels que soient $\tau_i > 0$, $i = 1, \ldots, s$.

Le problème

$$q(\tau_1^0, \ldots, \tau_s^0) = \min_{\tau_1, \ldots, \tau_s} q(\tau_1, \ldots, \tau_s) \tag{4.39}$$

a été résolu pour la première fois par Zolotarev [IX; 6]. Peaceman, Rachford [XV], Douglas, Rachford [XV], Jordan (voir Wachspress [XV; 38d)]), Wachspress [XV] l'ont abordé dans le cas commutatif. On établit en particulier à l'aide des résultats de ces auteurs qu'on peut choisir la longueur s du cycle et les paramètres $\check{\tau} = (\check{\tau}_1, \ldots, \check{\tau}_s)$, solution approchée de (4.39), de façon qu'on obtienne pour A mal conditionnée

$$\frac{C}{\ln p} \leqslant R(T_{\check{\tau}}^{(s)}),$$

où $p = p(A)$, C étant une constante indépendante de s et p. Autrement dit, cette méthode l'emporte en vitesse de convergence sur toutes les techniques considérées.

Bornons-nous au schéma à un paramètre (4.21) et essayons de résoudre le problème d'optimisation d'une autre manière. Il découle des résultats précédents que, quel que soit $\tau > 0$, la méthode de décomposition amortit à chaque itération tout coefficient ψ_n^j du développement du vecteur erreur

$$\psi^j = \sum_n \psi_n^j u_n$$

et qu'avec τ optimal, on affaiblit au mieux les coefficients, et ce uniformément par rapport à toutes les valeurs de n. Supposons que les coefficients $\{\psi_n^0\}$ du développement du vecteur erreur initiale ψ^0 sont prépondérants pour les valeurs propres $\lambda_n \in [\widetilde{m}, \Delta]$, $\widetilde{m} = \min(\alpha(A_1), \alpha(A_2)) < \Delta$, $\Delta \ll \widetilde{M} = \max(\beta(A_1), \beta(A_2))$. Il y a visiblement intérêt à effectuer plusieurs itérations (4.21) pour τ

choisi par la formule

$$\tau = \frac{1}{\sqrt{\widetilde{m}\Delta}}.$$

Avec ce τ, on atténue au mieux et uniformément les coefficients prépondérants au bout de plusieurs premières itérations, puis on reprend τ défini par (4.29).

Dans la pratique, c'est les coefficients du développement suivant les vecteurs propres associés aux plus petites valeurs propres qui s'avèrent prépondérants, et on admet souvent (en idéalisant légèrement) qu'ils sont monotones décroissants avec l'augmentation de λ_n. Physiquement, cela correspond à la prépondérance des processus à grande échelle (λ_n petites) par rapport à ceux à échelle peu importante (λ_n grandes). On choisit évidemment Δ à la lumière des renseignements à priori concrets sur le problème.

4.4.2. Cas non commutatif

Considérons la méthode de décomposition

$$(E + \tau A_1)(E + \tau A_2)(\varphi^{j+1} - \varphi^j) = -2\tau(A\varphi^j - f) \qquad (4.40)$$

à A_1, A_2 définies positives qui ne commutent pas. On introduit la matrice définie positive symétrique

$$D_\tau = D_{1,\tau}^* D_{1,\tau} \qquad (4.41)$$

et les vecteurs

$$z^j = D_{1,\tau}\psi^j. \qquad (4.42)$$

Ici

$$D_{1,\tau} = (E + \tau A_1)^{-1} A \qquad (4.43)$$

et $\psi^j = \varphi^j - \varphi^*$ sont les vecteurs erreurs. Evidemment, $z^j = (E + \tau A_1)^{-1}\xi^j$, où $\xi^j = A\varphi^j - f$ sont les vecteurs résidus.

On montre par des transformations standard que

$$\psi^{j+1} = T_\tau \psi^j, \qquad (4.44)$$

où

$$T_\tau = E - 2\tau(E + \tau A_2)^{-1}(E + \tau A_1)^{-1} A \qquad (4.45)$$

est l'opérateur de transition de (4.40), et que

$$z^{j+1} = \widetilde{T}_\tau z^j, \qquad (4.46)$$

où

$$\widetilde{T}_\tau = T_{1,\tau} T_{2,\tau} \qquad (4.47)$$

et

$$T_{i,\tau} = (E - \tau A_i)(E + \tau A_i)^{-1}, \quad i = 1, 2. \qquad (4.48)$$

On a établi (4.46) à (4.48) à l'aide des relations

$$\begin{aligned}\widetilde{T}_\tau &= (E+\tau A_1)^{-1} A T_\tau A^{-1} (E+\tau A_1) = \\ &= (E+\tau A_1)^{-1} [(E+\tau A_1)(E+\tau A_2) - 2\tau A] (E+\tau A_2)^{-1} = \\ &= (E+\tau A_1)^{-1} (E-\tau A_1)(E-\tau A_2)(E+\tau A_2)^{-1} = \\ &= (E-\tau A_1)(E+\tau A_1)^{-1} (E-\tau A_2)(E+\tau A_2)^{-1}.\end{aligned}$$

Ainsi, tout vecteur ψ ($z = D_{1,\tau}\psi$, $x = T_{2,\tau} z$) vérifie

$$\| T_\tau \psi \|_{D_\tau} = \| \widetilde{T}_\tau z \|_2 = \frac{\| T_{1,\tau} x \|_2}{\| x \|_2} \frac{\| T_{2,\tau} z \|_2}{\| z \|_2} \| \psi \|_{D_\tau}, \qquad (4.49)$$

donc

$$\| T_\tau \|_{D_\tau} = \| \widetilde{T}_\tau \|_2 \leqslant \| T_{1,\tau} \|_2 \| T_{2,\tau} \|_2. \qquad (4.50)$$

Etudions $\| T_{i,\tau} \|_2$, $i = 1, 2$, sous l'hypothèse de validité des inégalités

$$\begin{aligned}(A_i u, u) &\geqslant m (u, u), \\ (A_i u, A_i u) &\leqslant M (A_i u, u), \qquad i = 1, 2,\end{aligned} \qquad (4.51)$$

quel que soit u.

Par définition,

$$\begin{aligned}\| T_{i,\tau} \|_2^2 &= \sup_{v \neq 0} \frac{\| T_{i,\tau} v \|_2^2}{\| v \|_2^2} = \sup_{u \neq 0} \frac{\| (E-\tau_i A_i) u \|_2^2}{\| (E+\tau_i A_i) u \|_2^2} = \\ &= \sup_{u \neq 0} \frac{(u, u) - 2\tau (A_i u, u) + \tau^2 (A_i u, A_i u)}{(u, u) + 2\tau (A_i u, u) + \tau^2 (A_i u, A_i u)} = \\ &= 1 - 4\tau \inf_{u \neq 0} \frac{(A_i u, u)}{(u, u) + 2\tau (A_i u, u) + \tau^2 (A_i u, A_i u)}, \end{aligned} \qquad (4.52)$$

avec $u = (E + \tau A_i)^{-1} v$. Les relations ci-dessus entraînent que le rapport

$$\frac{\| T_{i,\tau} v \|_2}{\| v \|_2}$$

est soit supérieur à 1 pour $\tau < 0$, soit inférieur à 1 ($\tau > 0$) quels que soient le vecteur v non nul et i. Selon (4.49), une condition nécessaire de convergence de la méthode de décomposition (4.40) est donc la positivité de τ. Dans la suite, τ est partout supposé > 0.

On évalue $\| T_{i,\tau} \|_2$ moyennant (4.51) et (4.52):

$$\begin{aligned}\| T_{i,\tau} \|_2^2 &\leqslant 1 - 4\tau \inf_{u \neq 0} \frac{(A_i u, u)}{(u, u) + 2\tau (A_i u, u) + \tau^2 M (A_i u, u)} = \\ &= 1 - 4\tau \inf_{\|w\|_2 = 1} \frac{(A_i w, w)}{1 + (2\tau + M\tau^2)(A_i w, w)} \leqslant 1 - 4\tau \inf_{t \geqslant m} \frac{t}{1 + (2\tau + M\tau^2) t} = \\ &= 1 - 4\tau \frac{m}{1 + (2\tau + M\tau^2) m} = \frac{1 - 2m\tau + mM\tau^2}{1 + 2m\tau + mM\tau^2} < 1\end{aligned}$$

à l'aide des notations $w = \frac{u}{\|u\|_2}$, $t = (A_i w, w)$ et de la croissance monotone de la fonction

$$u(t) = \frac{t}{1+at}$$

pour $t \geqslant 0$ et tout a positif. Ainsi, sous les hypothèses (4.51), on a

$$\|\widetilde{T}_\tau\|_{D_\tau} \leqslant \|T_{1,\tau}\|_2 \, \|T_{2,\tau}\|_2 \leqslant \frac{1-2m\tau+mM\tau^2}{1+2m\tau+mM\tau^2} < 1$$

quel que soit $\tau > 0$.

Le problème d'optimisation approchée de la méthode (4.40) exige qu'on minimise

$$q(\tau) = \frac{1-2m\tau+mM\tau^2}{1+2m\tau+mM\tau^2} = 1 - \frac{4m\tau}{1+2m\tau+mM\tau^2} \tag{4.53}$$

par rapport à τ. Comme $g(0) = g(+\infty) = 1$ et $q(\tau)$ admet pour $\tau \in (0, +\infty)$ des dérivées de tout ordre et $0 \leqslant q(\tau) < 1$, la valeur $\tau = \tau_{\text{opt}}$ minimisant $q(\tau)$ est solution de l'équation

$$\frac{dq(\tau)}{d\tau} \equiv -4m \frac{1-mM\tau^2}{(1+2m\tau+mM\tau^2)^2} = 0.$$

D'où

$$\tau_{\text{opt}} = \frac{1}{\sqrt{mM}}, \tag{4.54}$$

et

$$q(\tau_{\text{opt}}) = \frac{1-\sqrt{\frac{m}{M}}}{1+\sqrt{\frac{m}{M}}}, \tag{4.55}$$

$$R(T_{\tau_{\text{opt}}}) \geqslant -\ln\left[\frac{1-\sqrt{\frac{m}{M}}}{1+\sqrt{\frac{m}{M}}}\right]. \tag{4.56}$$

Voyons ce que représente cette estimation. Supposons A_1, A_2 symétriques, $\alpha(A_1) = \alpha(A_2) = \frac{1}{2}\alpha(A)$ et $\beta(A_1) = \beta(A_2) = \frac{1}{2}\beta(A)$. On voit facilement que

$$m = \frac{1}{2}\alpha(A), \quad M = \frac{1}{2}\beta(A),$$

donc

$$q(\tau_{\text{opt}}) = \frac{\sqrt{p}-1}{\sqrt{p}+1}.$$

Ainsi, on a pour les systèmes mal conditionnés ($p = p(A) \gg 1$)

$$R(T_{\tau_{opt}}) \geqslant \frac{2}{\sqrt{p}}. \tag{4.57}$$

D'autre part, l'analyse spectrale directe fournit pour $\tau = \tau_{opt}$ (comme au point 4.4.1)

$$\| T_{i,\tau_{opt}} \|_2 = \max_{m \leqslant \lambda \leqslant M} \left| \frac{1-\tau_{opt}\lambda}{1+\tau_{opt}\lambda} \right| = \frac{1-\sqrt{\frac{m}{M}}}{1+\sqrt{\frac{m}{M}}}, \tag{4.58}$$

donc

$$\beta(T_{\tau_{opt}}) \leqslant \left[\frac{1-\sqrt{\frac{m}{M}}}{1+\sqrt{\frac{m}{M}}} \right]^2 \leqslant \left[\frac{\sqrt{p}-1}{\sqrt{p}+1} \right]^2. \tag{4.59}$$

S'agissant des systèmes mal conditionnés,

$$\begin{aligned} \beta(T_{\tau_{opt}}) &\leqslant 1 - \frac{4}{\sqrt{p}}, \\ R(T_{\tau_{opt}}) &\geqslant \frac{4}{\sqrt{p}}, \end{aligned} \tag{4.60}$$

i.e. lorsque $\tau = \tau_{opt}$, le processus (4.40) converge au moins aussi vite que chacune des méthodes itératives stationnaires considérées. Il y a lieu de noter la différence entre (4.57) et (4.60) dont la cause est le caractère plus général du premier procédé d'évaluation. Néanmoins on améliore sensiblement les estimations en recourant aux propriétés auxiliaires des matrices A_1, A_2.

Quelques mots enfin sur une sous-classe importante de schémas de la méthode (4.40). Soit le système $A\varphi = f$ à A supposée symétrique et définie positive,

$$\begin{aligned} A &= A_1 + A_2, \\ A_1 &= A_2^*. \end{aligned} \tag{4.61}$$

Etant donnée la réalité de A_1, A_2, tout ψ de l'espace des vecteurs réels vérifie sous les conditions (4.61) les égalités

$$(A_1\psi, \psi) = (A_2\psi, \psi) = \frac{1}{2}(A\psi, \psi)$$

Il en résulte que A_1, A_2 sont définies positives dès que A l'est, et comme déjà vu antérieurement, la méthode de décomposition (4.40) à ces matrices converge pour tout $\tau > 0$.

On construit d'ordinaire A_1 et A_2 moyennant A donnée de façon suivante. On décompose A en blocs:

$$A = \left\| \begin{matrix} A_{11} & A_{12} & \dots & A_{1k} \\ A_{21} & A_{22} & \dots & A_{2k} \\ \dots & \dots & \dots & \dots \\ A_{k1} & A_{k2} & \dots & A_{kk} \end{matrix} \right\|$$

à A_{ii} diagonales carrées. On prend pour A_1 une matrice tridiagonale par blocs de la forme

$$A_1 = \left\| \begin{matrix} \frac{1}{2}A_{11} & 0 & & 0 \\ A_{21} & \frac{1}{2}A_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ A_{k1} & A_{k2} & \dots & \frac{1}{2}A_{kk} \end{matrix} \right\|$$

et on suppose $A_2 = A_1^*$. Un cas particulier simple est $A_{ii} = a_{ii}$, i.e. A_{ii} sont égales aux éléments diagonaux de A, auquel cas A_1 est triangulaire inférieure et A_2 triangulaire supérieure. Les méthodes de décomposition avec $A_1 = A_2^*$ ont été étudiées par Samarski [XV; 11] et Iline [XV; 7].

4.4.3. Optimisation des méthodes de décomposition au sens de Tchebycheff et du Calcul des variations

Il est déjà connu que dans la méthode de décomposition

$$(E + \tau A_1)(E + \tau A_2)(\varphi^{j+1} - \varphi^j) = -\gamma_j (A\varphi^j - f), \quad (4.62)$$

$\gamma_j = 2\tau$, on a les inégalités

$$q_j = \frac{\| (E - \gamma_j H_\tau A)\psi^j \|_{D_\tau}}{\| \psi^j \|_{D_\tau}} \leqslant \| T_\tau \|_{D_\tau} < 1, \quad (4.63)$$

$\tau > 0$ quelconque, où

$$H_\tau = (E + \tau A_2)^{-1}(E + \tau A_1)^{-1},$$
$$T_\tau = E - 2\tau H_\tau A$$

et ψ^j est le vecteur erreur. Selon le point 4.3.1, cela signifie la convergence de la méthode itérative non stationnaire (4.62) avec paramètres γ_j, solution de l'équation

$$\frac{d}{d\gamma_j} \| (E - \gamma_j H_\tau A)\psi^j \|_{D_\tau} = 0, \quad (4.64)$$

i.e. choisis à partir de la condition de minimum à chaque pas de la fonctionnelle $J(\varphi) = (D_\tau(\varphi^* - \varphi), \varphi^* - \varphi) = \| \varphi^* - \varphi \|^2_{D_\tau}$. Ceci étant, la norme de l'opérateur de transition $T_{\tau, \gamma}$ est majorée pour tout $\tau > 0$ par

$$\| T_{\tau, \gamma} \|_{D_\tau} \leqslant \| T_\tau \|_{D_\tau} < 1. \tag{4.65}$$

On résout (4.64) et on trouve pour γ_j :

$$\gamma_j = \frac{(H_\tau A \psi^j, \psi^j)_{D_\tau}}{\| H_\tau A \psi^j \|^2_{D_\tau}} = \frac{((E + \tau A_1)^{-1} A H_\tau \xi^j, (E + \tau A_1)^{-1} \xi^j)}{\| (E + \tau A_1)^{-1} A H_\tau \xi^j \|^2_2}, \tag{4.66}$$

$\xi^j = A\varphi^j - f$ étant le vecteur résidu.

On désire maintenant optimiser (4.40) sous l'hypothèse complémentaire $H_\tau = H^*_\tau$. Il découle des points précédents de ce numéro qu'on est alors ou bien dans le cas commutatif si $A_i = A^*_i$, $i = 1, 2$, ou bien dans le cas $A_1 = A^*_2$ où

$$H_\tau = (E + \tau A_2)^{-1} (E + \tau A_1)^{-1} = (E + \tau A_1^*)^{-1} (E + \tau A_1)^{-1} =$$
$$= [(E + \tau A_1)^{-1}]^* (E + \tau A_1)^{-1} = H^*_\tau.$$

Fort de l'hypothèse faite et des résultats du point 4.2.4, on conclut à la possibilité de choisir les paramètres $\{\gamma_j\}$ de la méthode itérative s-cyclique

$$\varphi^{j+1} = \varphi^j - \gamma_j H_\tau (A\varphi^j - f) \tag{4.67}$$

au sens de la théorie générale des méthodes itératives cycliques de Tchebycheff (voir point 4.2.4). Les quantités $\alpha = \alpha(A)$ et $\beta = \beta(A)$ font place respectivement à

$$m = \alpha(H_\tau A)$$

et à

$$M = \beta(H_\tau A).$$

L'évaluation ou le calcul de m et M utilisent les estimations de la norme de l'opérateur $T_\tau = E - 2\tau H_\tau A$ ou des processus itératifs auxiliaires du type Ljusternik.

Plaçons-nous par exemple dans le cas commutatif et évaluons m et M si $\tau = \tau_{\text{opt}}$ (voir (4.29)) et A est une matrice mal conditionnée. La première inégalité (4.32) entraîne

$$m = \alpha(H_\tau A) \geqslant \frac{1}{2\tau} \frac{4}{\sqrt{p(A)}},$$
$$M = \beta(H_\tau A) \leqslant \frac{1}{2\tau} \left(2 - \frac{4}{\sqrt{p(A)}}\right), \tag{4.68}$$

et on aboutit à

$$p(H_\tau A) = \frac{\beta(H_\tau A)}{\alpha(H_\tau A)} \geqslant \frac{1}{2}(\sqrt{p(A)} - 2) \approx \frac{\sqrt{p(A)}}{2}. \tag{4.69}$$

Если учесть (2.69) и то, что $p(H_\tau A) \gg 1$ при $p(A) \gg 1$, получим для метода (4.67) при $s \gg 1$:

$$R(T_{\gamma,\tau}) \approx \frac{2}{\sqrt{p(H_\tau A)}} \approx \frac{2\sqrt{2}}{\sqrt[4]{p(A)}}. \tag{4.70}$$

Таким образом, применение чебышевского ускорения увеличивает в $\sqrt{\frac{p(A)}{2}}$ раз эффективность метода декомпозиции.

Если $A_1 = A_2^*$, то сходимость ускоряется методом сопряженных градиентов. Поскольку H_τ симметрична и положительно определена, формулы (3.34) принимают вид

$$\begin{aligned} g_k &= \begin{cases} H_\tau \xi^0, & k=1, \\ H_\tau \xi^{k-1} - b_k g_{k-1}, & k>1, \end{cases} \\ b_k &= \frac{(AH_\tau \xi^{k-1},\ g_{k-1})}{(Ag_{k-1},\ g_{k-1})}, \\ \varphi^k &= \varphi^{k-1} - \alpha_k g_k, \\ \alpha_k &= \frac{(\xi^{k-1},\ g_k)}{(Ag_k,\ g_k)}, \\ k &= 1, \ldots, s. \end{aligned} \tag{4.71}$$

Отметим, что скорость сходимости этого метода асимптотически не меньше, чем у (4.67) с $\{\gamma_j\}$, выбранными для данного s.

4.5. Итерационные методы для систем с вырожденными матрицами

Рассмотрим систему линейных алгебраических уравнений

$$A\varphi = f \tag{5.1}$$

с симметричной и полуопределенной матрицей A. Обозначим через $\{u_n\}$ ортонормированную систему собственных векторов A, соответствующих собственным значениям $\{\lambda_n\}$, и через $\ker A$ — нулевое подпространство A, т. е. множество векторов ψ, для которых $A\psi = 0$. Предположим, что $\ker A$ имеет размерность m и что ненулевые собственные значения суть $\{\lambda_n\}_{n=1}^m$ (так что $u_n \in \ker A$ при $n = 1, \ldots, m$).

Разложим векторы f и φ из (5.1) по базису $\{u_n\}$:

$$f = \sum_n f_n u_n,$$

$$\varphi = \sum_n \varphi_n u_n$$

и подставим в (5.1), получим

$$\lambda_n \varphi_n = f_n$$

и, в частности,

$$\lambda_n \varphi_n = f_n \text{ при } n = 1, \ldots, m.$$

Les dernières égalités ne sont possibles que pour $f_n = 0$, $n = 1, \ldots, m$, si bien que le système est compatible sous la condition d'orthogonalité $f \perp \ker A$ ou, chose équivalente,

$$(f, u_n) = 0, \qquad n = 1, \ldots, m.$$

S'agissant de (5.1) incompatible, on prend souvent comme sa solution la solution φ^* du système compatible correspondant

$$A\varphi = \tilde{f}, \tag{5.2}$$

où

$$\tilde{f} = \sum_{n>m} f_n u_n,$$

et on dit que φ^* est *solution généralisée* de (5.1). Les solutions généralisées minimisent, on le montre facilement, la norme euclidienne du vecteur résidu. Si la matrice A de (5.1) est quelconque, le vecteur φ^* constitue une solution généralisée de ce système s'il est solution du problème

$$\| A\varphi^* - f \|_2 = \min_{\varphi} \| A\varphi - f \|_2$$

(voir p.ex. Voévodine [VIII; 3], Faddeev, Faddeeva, Koublanovskaïa [VIII; 17]).

4.5.1. Un système compatible

Supposons le système (5.1) compatible et appliquons la méthode itérative stationnaire

$$B(\varphi^{j+1} - \varphi^j) = -(A\varphi^j - f) \tag{5.3}$$

à B symétrique et définie positive. Quelle est la propriété de convergence du processus? On introduit les vecteurs $\psi^j = \varphi^j - \varphi^*$, φ^* étant une solution quelconque mais fixe de (5.1), et les itérations pour $\{\psi^j\}$ s'écrivent

$$\psi^{j+1} = T\psi^j, \tag{5.4}$$

où

$$T = E - B^{-1}A \tag{5.5}$$

désigne l'opérateur de transition de (5.3). On introduit les vecteurs $z^j = B^{1/2}\psi^j$ et on obtient en multipliant (5.4) par $B^{1/2}$

$$z^{j+1} = \tilde{T}z^j, \tag{5.6}$$

avec

$$\tilde{T} = E - B^{-1/2}AB^{-1/2} \tag{5.7}$$

une matrice symétrique, la matrice $S = B^{-1/2}AB^{-1/2}$ étant symétrique et semi-définie positive (ici $B^{-1/2} = [B^{1/2}]^{-1}$).

On note $\{v_n\}$ le système orthonormé complet de vecteurs propres associés aux valeurs propres $\{\mu_n\}$ de $\tilde{T}$ et on développe z^j par rap-

port à ce système:

$$z^j = \sum_n z_n^j v_n. \tag{5.8}$$

Le report de (5.8) dans (5.6) donne les relations

$$z_n^{j+1} = \mu_n z_n^j = \mu_n^{j+1} z_n^0.$$

D'où la nécessité des conditions

$$\begin{cases} |\mu_n| \leqslant 1, \\ \mu_n \neq -1 \end{cases} \tag{5.9}$$

pour que la méthode (5.3) converge pour toute approximation initiale φ^0. En effet, la suite des z_n^j est divergente ou ne converge vers aucune quantité concrète pour tout $z_n^0 \neq 0$ selon que $|\mu_n| > 1$ ou $\mu_n = -1$.

On démontre que (5.9) sont des conditions suffisantes pour que la suite $\{\varphi^j\}$ converge vers une solution $\widetilde{\varphi}^*$ de (5.1) quelle que soit l'approximation initiale φ^0. L'inégalité $|\mu_n| < 1$ garantit évidemment la convergence de z_n^j vers 0 quand $j \to \infty$ et on a dans le cas $\mu_n = 1$

$$z_n^{j+1} = z_n^j.$$

Supposons qu'on est dans les conditions (5.9), que $\mu_n = 1$ est d'ordre de multiplicité s (on démontre que $s = m$) et que $\mu_1 = \mu_2 = \ldots = \mu_s = 1$. On vérifie sans peine que

$$\lim_{j\to\infty} z^j = z^\infty = \sum_{n=1}^{s} z_n^0 v_n.$$

L'égalité $Tz^\infty = z^\infty$ ($\widetilde{T}v_n = v_n$ pour $n = 1, \ldots, s$) entraîne

$$B^{-1/2} A B^{-1/2} z^\infty = 0,$$

donc

$$B^{-1/2} z^\infty \in \ker A. \tag{5.10}$$

D'autre part,

$$B^{-1/2} z^\infty = \lim_{j\to\infty} B^{-1/2} z^j = \lim_{j\to\infty} \psi^j = \psi^\infty. \tag{5.11}$$

On conclut par regroupement de (5.10) et (5.11) que la suite $\{\varphi^j\}$ définie par (5.3) a pour limite un vecteur

$$\varphi^\infty = \varphi^* + \psi^\infty \tag{5.12}$$

qui est solution du système (5.1) en vertu de l'appartenance de ψ^∞ à ker A, i.e. $\varphi^\infty = \widetilde{\varphi}^*$ et dépend de l'approximation initiale φ^0.

Les conditions suffisantes de convergence des méthodes itératives générales stationnaires ou non pour les systèmes à matrices dégénérées ont été étudiées par Kouznétsov (voir Marchouk et Kouznétsov [VIII ; 12]).

4.5.2. Un système incompatible

S'agissant de (5.1) incompatible, on l'aborde de deux manières. 1. Supposons les matrices B et A de (5.3) posséder un même système orthonormé complet de vecteurs propres $\{u_n\}$, i.e.

$$\begin{aligned} Bu_n &= \nu_n u_n, \\ Au_n &= \lambda_n u_n, \end{aligned} \tag{5.13}$$

et la méthode vérifier les conditions (5.9). La dernière hypothèse signifie, on vient de l'établir, la convergence de (5.3) dans le cas de systèmes compatibles correspondants. Décomposons les vecteurs φ^j et f dans la base $\{u_n\}$:

$$\varphi^j = \sum_n \varphi_n^j u_n,$$

$$f = \sum_n f_n u_n,$$

il vient

$$\varphi_n^{j+1} = \left(1 - \frac{\lambda_n}{\nu_n}\right) \varphi_n^j + \frac{1}{\nu_n} f_n = \left(1 - \frac{\lambda_n}{\nu_n}\right)^{j+1} \varphi_n^0 + \frac{1}{\nu_n} \sum_{k=0}^{j} \left(1 - \frac{\lambda_n}{\nu_n}\right)^k f_n.$$

D'où, compte tenu des conditions (5.9),

$$\lim_{j \to \infty} \varphi_n^j = \frac{1}{\lambda_n} f_n$$

pour $\lambda_n \neq 0$, $n > m$, et

$$\varphi_n^j = \frac{j}{\nu_n} f_n + \varphi_n^0$$

pour $\lambda_n = 0$, $n \leqslant m$. Ainsi, il y a divergence de la suite φ^j si $f_n \neq 0$ pour au moins un $n \leqslant m$.

Formons la suite de

$$z^j = \varphi^j - j\,(\varphi^{j+1} - \varphi^j) \tag{5.14}$$

et décomposons les vecteurs z^j dans la base $\{u_n\}$:

$$z^j = \sum_n z_n^j u_n.$$

On a

$$z_n^j = \begin{cases} \varphi_n^j - j\left[-\dfrac{\lambda_n}{\nu_n}\left(1 - \dfrac{\lambda_n}{\nu_n}\right)^j \varphi_n^0 + \dfrac{1}{\nu_n}\left(1 - \dfrac{\lambda_n}{\nu_n}\right)^j f_n\right], & n > m, \\ \varphi_n^0, & n \leqslant m, \end{cases}$$

donc

$$\lim_{j\to\infty} z_n^j = \begin{cases} \frac{1}{\lambda_n} f_n, & n > m, \\ \varphi_n^0, & n \leqslant m. \end{cases} \tag{5.15}$$

Ainsi, le vecteur

$$z^\infty = \lim_{j\to\infty} z^j \tag{5.16}$$

existe et est solution du système compatible (5.2).

Des méthodes itératives (5.3), (5.14) pour les systèmes incompatibles d'équations linéaires à matrices symétriques et semi-définies positives ont été analysées par Moltchanov et Yakovlev [VIII; 13].

2. Supposons que la dimension de l'espace nul de la matrice A est 1 ($m = 1$), que B du processus itératif (5.3) est symétrique et définie positive et qu'on est dans les conditions (5.9). Il est évident qu'il y a coïncidence des espaces nuls des A et $B^{-1}A$, que $B^{-1}A$ a toutes ses valeurs propres réelles et qu'elle possède un système complet de vecteurs propres (voir point 4.2.4. pour la dernière propriété).

Notons $\{v_n\}$ le système de vecteurs propres de $B^{-1}A$ et $\{v_n^*\}$ celui de $(B^{-1}A)^* = AB^{-1}$. S'ils sont complets, ils constituent une base biorthogonale dans l'espace primitif des vecteurs, i.e.

$$(v_i, v_j^*) = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

On estime que $v_1 \in \ker (B^{-1}A) = \ker A$. Développons les vecteurs x^j de la méthode itérative

$$\begin{aligned} x^{j+1} &= x^j - B^{-1}Ax^j, \\ x^0 &= B^{-1}f \end{aligned} \tag{5.17}$$

par rapport au système $\{v_n\}$:

$$x^j = \sum_n x_n^j v_n, \qquad x_n^j = (x^j, v_n^*).$$

On a, conformément à (5.9),

$$x^\infty = x_1^0 v_1.$$

Cela signifie avec l'hypothèse $x_1^0 = (B^{-1}f, v_1^*) = 0$ la compatibilité de

$$B^{-1}A\varphi = B^{-1}f$$

et, partant, la même propriété du système initial: contradiction. Ainsi, $x_1^0 \neq 0$ et $v^\infty = x_1^0 v_1 \in \ker A$. Rendons f orthogonal à v^∞:

$$\tilde{f} = f - \frac{(f, v^\infty)}{(v^\infty, v^\infty)} v^\infty, \tag{5.18}$$

il vient le système

$$A\varphi = \tilde{f} \tag{5.19}$$

compatible et équivalent au système initial au sens de l'ensemble des solutions généralisées.

La méthode a été proposée par Kouznétsov qui l'a généralisée au cas de matrices non symétriques et d'espaces nuls de dimension arbitraire (voir [VIII ; 9, 28]).

4.5.3. Analogue matriciel de la méthode du domaine auxiliaire

Soit le système algébrique linéaire

$$A_0\varphi_0 = f_0 \tag{5.20}$$

à matrice A_0 symétrique et définie positive. Considérons de plus le système

$$A\varphi = f \tag{5.21}$$

où

$$A = \begin{Vmatrix} A_0 & 0 \\ 0 & A_1 \end{Vmatrix}, \tag{5.22}$$

A_1 étant symétrique et semi-définie positive, $\varphi = \begin{Vmatrix} \varphi_0 \\ \varphi_1 \end{Vmatrix}$ et $f = \begin{Vmatrix} f_0 \\ 0 \end{Vmatrix}$. La matrice A est évidemment symétrique et semi-définie positive et le vecteur φ_0 de toute solution

$$\varphi = \begin{Vmatrix} \varphi_0 \\ \varphi_1 \end{Vmatrix}$$

du système (5.21) est solution de (5.20).

On résout (5.21) par la méthode itérative stationnaire

$$\varphi^{j+1} = \varphi^j - H(A\varphi^j - f) \tag{5.23}$$

à matrice H symétrique et définie positive, dont l'opérateur de transition

$$T = E - HA \tag{5.24}$$

a toutes ses valeurs propres vérifiant la condition (5.9). On a montré au point 4.5.1. que s'agissant de (5.21) compatible (or, il l'est toujours), la vitesse de convergence du procédé considéré est définie par la valeur propre de plus grand module non égal à 1 de T. Kouznétsov et Matsokine [IV ; 11] ont montré qu'en ce qui concerne la minimisation de cette quantité, le choix optimal est

$$A_1 = 0$$

pour toute H définie positive symétrique.

Prenons à titre d'exemple le système d'équations à cinq points qui approchent le problème de Dirichlet relatif à l'équation de Poisson dans le domaine en L de la fig. 16. Les petites croix repèrent ici les nœuds où l'on se donne les équations compte tenu des conditions aux limites et les cercles blancs jouent le même rôle pour les équations à coefficients nuls (ce qui correspond à $A_1 = 0$).

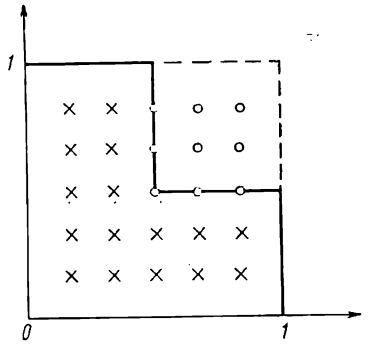


Fig. 16

Plaçons-nous dans le cas où l'on prend comme $B = \tau H^{-1}$ le laplacien discrétisé à cinq points basé sur tous les nœuds d'un réseau rectangulaire de pas h qui recouvrent le carré. On a établi la majoration pour la plus grande valeur propre ($\neq 1$) de T ($h \ll 1$)

$$|\lambda_{\max}| \leqslant 1 - Ch, \qquad (5.25)$$

C étant une constante indépendante de h.

On démontre l'estimation (5.25). Soit $h = 1/(n+1)$, $k = (n+1)/2$. On numérote les nœuds en commençant par ceux du domaine en L, on écrit les équations à cinq points correspondantes

$$\frac{-u_{i-1j}+2u_{ij}-u_{i+1j}}{h^2}+\frac{-u_{ij-1}+2u_{ij}-u_{ij+1}}{h^2}=f_{ij} \qquad (5.26)$$

et on adjoint les équations à coefficients nuls, il vient le problème matriciel

$$Au = f$$

de la forme (5.21), (5.22) à A d'ordre n^2 :

$$A=\begin{Vmatrix} A_0 & 0 \\ 0 & 0 \end{Vmatrix}.$$

La matrice du laplacien discrétisé s'écrit dans ce cas

$$B=\begin{Vmatrix} A_0 & -C \\ -C^T & B_1 \end{Vmatrix},$$

B_1 étant symétrique et définie positive.

On trouve la plus petite valeur propre non nulle de la matrice HA à partir de l'équation

$$\tau A\varphi = \lambda B\varphi, \qquad (5.27)$$

avec $\varphi = (\varphi_0, \varphi_1)^T$ vecteur propre associé. On en tire facilement l'égalité

$$\varphi_1 = B_1^{-1}C^T\varphi_0. \qquad (5.28)$$

Multiplions (5.27) scalairement par φ, il vient, compte tenu de (5.28),

$$\lambda = \tau \frac{(A\varphi,\ \varphi)}{(B\varphi,\ \varphi)} = \tau \frac{(A_0\varphi_0,\ \varphi_0)}{(A_0\varphi_0,\ \varphi_0) - (B_1^{-1}C^T\varphi_0,\ C^T\varphi_0)} \geqslant \tau \qquad (5.29)$$

parce que $(B\varphi,\ \varphi) > 0$ et $(B_1^{-1}C^T\varphi_0,\ C^T\varphi_0) > 0$.

Supposons que $\lambda = \rho$ est la plus grande valeur propre du problème (5.27) et φ un vecteur propre associé. On introduit les matrices carrées d'ordre n

$$A_n = \frac{1}{h^2}\begin{Vmatrix} 2 & -1 & 0 \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 \ldots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 \ldots & 0 & -1 & 2 \end{Vmatrix}, \qquad \hat{A}_n = \begin{Vmatrix} A_{k-1} & 0 \\ 0 & 0 \end{Vmatrix}$$

et les vecteurs de dimension n dont les composantes sont celles de φ :

$$\varphi^i = \begin{Vmatrix} \varphi_{i1} \\ \varphi_{i2} \\ \vdots \\ \varphi_{in} \end{Vmatrix}, \quad i = 1, \ldots, n, \qquad \psi^j = \begin{Vmatrix} \varphi_{1j} \\ \varphi_{2j} \\ \vdots \\ \varphi_{nj} \end{Vmatrix}, \quad j = 1, \ldots, n.$$

On multiplie (5.27) scalairement par φ :

$$\rho = \frac{(A\varphi,\ \varphi)}{(H^{-1}\varphi,\ \varphi)} = \tau \frac{(A\varphi,\ \varphi)}{(B\varphi,\ \varphi)} =$$

$$= \tau \frac{\sum\limits_{i=1}^{k-1} [(A_n\varphi^i,\ \varphi^i) + (A_n\psi^i,\ \psi^i)] + \sum\limits_{i=1}^{n} [(\hat{A}_n\varphi^i,\ \varphi^i) + (\hat{A}_n\psi^i,\ \psi^i)]}{\sum\limits_{i=1}^{n} [(A_n\varphi^i,\ \varphi^i) + (A_n\psi^i,\ \psi^i)]}.$$

Si un produit scalaire du dénominateur s'annule, il en est de même du vecteur correspondant. L'inégalité

$$\frac{\sum\limits_{i=1}^{m} a_i}{\sum\limits_{i=1}^{m} b_i} \leqslant \max_{1 \leqslant i \leqslant m} \frac{a_i}{b_i}$$

est vérifiée pour tout $a_i \geqslant 0$ et tout $b_i > 0$, si bien qu'elle conduit à

$$\rho \leqslant \tau \max\left\{1,\ \frac{(\hat{A}_n g,\ g)}{(A_n g,\ g)}\right\}, \qquad (5.30)$$

où $g \neq 0$ coïncide avec φ^i ou ψ^i, $i = 1, \ldots, n$.

La quantité $(\hat{A}_n g, g)/(A_n g, g)$ est au plus égale au rayon spectral de la matrice $G = A_n^{-1}\hat{A}_n$. Le calcul de G ne présente pas de difficulté :

$$G = \begin{Vmatrix} X & 0 \\ Y & 0 \end{Vmatrix},$$

où $X = (A_{k-1} - CA_k^{-1}C^T)^{-1} A_{k-1}$, $Y = A_k^{-1}C^T X$ et C est une matrice $(k-1) \times k$ dont tous les éléments sont nuls sauf l'élément en dernière ligne et première colonne, qui vaut $1/h^2$.

Il en découle la coïncidence du rayon spectral de G et de la plus grande valeur propre du problème

$$A_{k-1} v = \mu (A_{k-1} - CA_k^{-1}C^T) v$$

qu'on ramène facilement en explicitant $CA_k^{-1}C^T$ à

$$\begin{aligned}
\frac{1}{k+1}\varphi_{k-1} &= \left(1 - \frac{1}{\mu}\right)\varphi_1, \\
\frac{2}{k+1}\varphi_{k-1} &= \left(1 - \frac{1}{\mu}\right)\varphi_2, \\
&\cdots\cdots\cdots\cdots\cdots \\
\frac{k-1}{k+1}\varphi_{k-1} &= \left(1 - \frac{1}{\mu}\right)\varphi_{k-1},
\end{aligned}$$

d'où l'on tire de suite

$$\mu = \frac{k+1}{2} = h^{-1}\frac{1+2h}{4}.$$

Etant donné (5.30), on déduit

$$\rho \leqslant \frac{\tau}{h}\frac{1+2h}{4}. \tag{5.31}$$

Avec $\tau = 8h/(1 + 6h)$, on a l'estimation (5.25), où $C \approx 8$.

4.6. Méthodes itératives pour les problèmes à entrées inexactes

Soit l'équation opératorielle

$$A\varphi = f, \tag{6.1}$$

avec A une matrice définie positive et f un vecteur donné.

Lorsque nous avons eu à aborder des équations de la forme (6.1), A et f ont toujours été supposés donnés exactement. Dans la pratique, on a souvent affaire à des données initiales approchées, i.e. on a, au lieu de (6.1), l'équation

$$A^h \varphi^h = f^h, \tag{6.2}$$

où l'indice h repère l'imprécision due à toutes sortes d'erreurs (discrétisation, erreurs statistiques, erreurs aléatoires). Supposons connues les erreurs commises dans l'approximation des opérateurs et des vecteurs, i.e. on a des estimations à priori

$$\| (A - A^h)\, \varphi \| \leqslant \xi(h), \quad \| f - f^h \| \leqslant \eta(h). \tag{6.3}$$

Cherchons la solution du problème (6.1) moyennant l'équation (6.2) et l'information (6.3). Nous nous bornerons à un algorithme d'approximation simple car nos résultats se généralisent trivialement à la plupart des processus itératifs considérés. Soit donc la méthode itérative

$$[\varphi^h]^{j+1} = [\varphi^h]^j - \tau (A^h [\varphi^h]^j - f^h), \quad [\varphi^h]^0 = 0, \tag{6.4}$$

où le paramètre τ est supposé tel qu'on ait

$$q = \| E - \tau A^h \| < 1. \tag{6.5}$$

Quel est le critère d'arrêt de l'algorithme si l'on connaît à priori que les données initiales sont affectées d'erreurs du type (6.3)? Il est naturel qu'on poursuive les approximations jusqu'à ce que l'erreur de calcul au cours des itérations ne devienne approximativement égale à l'erreur de discrétisation. Si l'on obtient, après une itération, l'égalité de ces erreurs de nature différente, l'erreur sur la solution approchée ne se trouve jamais rattrapée. Il y a plus. S'agissant de A mal conditionnée, auquel cas $[A^h]^{-1}$ peut différer considérablement de A^{-1}, il se peut que les itérations (6.4) suivantes altèrent sérieusement le résultat. Un problème se pose : connaissant à priori les erreurs sur les données initiales, trouver le nombre optimal d'itérations au bout desquelles il y a concordance des erreurs. Ce problème a été étudié par Marchouk et Vassiliev [XVI; 12].

Revenons aux équations (6.1), (6.2) et résolvons-les formellement par rapport aux inconnues :

$$\varphi = A^{-1} f, \quad \varphi^h = [A^h]^{-1} f^h, \tag{6.6}$$

ce qui permet d'écrire

$$\varphi^h - \varphi = [A^h]^{-1} [f^h - f + (A - A^h)\, \varphi], \tag{6.7}$$

d'où

$$\| \varphi^h - \varphi \| \leqslant \| [A^h]^{-1} \| [\| f - f^h \| + \| (A - A^h)\, \varphi \|],$$

ou

$$\| \varphi^h - \varphi \| \leqslant \| [A^h]^{-1} \| [\xi(h) + \eta(h)] \tag{6.8}$$

si l'on prend en considération les informations à priori (6.3).

Etant donné le processus itératif (6.4), on obtient facilement

$$\varphi^h - [\varphi^h]^j = (E - \tau A^h)^j [A^h]^{-1} f^h,$$

d'où

$$\| \varphi^h - [\varphi^h]^j \| \leqslant q^j \| [A^h]^{-1} \| \| f^h \|. \tag{6.9}$$

Mais l'inégalité triangulaire entraîne

$$\| \varphi - [\varphi^h]^j \| \leqslant \| [\varphi^h]^j - \varphi^h \| + \| \varphi^h - \varphi \|,$$

d'où, par suite de (6.8), (6.9),

$$\| \varphi - [\varphi^h]^j \| \leqslant q^j \| [A^h]^{-1} \| \| f^h \| + \| [A^h]^{-1} \| [\xi(h) + \eta(h)]. \quad (6.10)$$

Le premier terme à droite évalue l'erreur commise dans le processus itératif et le second l'erreur due à l'imprécision des données. On exige que

$$q^j \| [A^h]^{-1} \| \| f^h \| = \| [A^h]^{-1} \| [\xi(h) + \eta(h)]. \quad (6.11)$$

C'est l'équation pour l'itération $j = j_0$ après laquelle on stoppe le processus :

$$j_0 = \frac{1}{\ln q} \ln \frac{\xi(h) + \eta(h)}{\| f^h \|}. \quad (6.12)$$

Comme j_0 est un entier naturel, on prend naturellement pour ce numéro la partie entière de l'expression correspondante.

On note que la formule (6.12) ne contient pas la norme de l'opérateur inverse, ce qui simplifie considérablement la recherche du nombre optimal cherché.

A part l'information à priori sur $\xi(h)$, $\eta(h)$ et $\| f^h \|$, la formule (6.12) renferme $q = \| E - \tau A^h \|$ qu'on trouve pour la norme euclidienne $\| \cdot \|_2$ à l'aide de la plus grande valeur propre de l'opérateur T^*T, où $T = E - \tau A^h$, i.e.

$$q = \sqrt{\beta(T^*T)}.$$

La borne supérieure du spectre de T est donnée par la méthode de Ljusternik du n° 1.1.

Les problèmes d'algèbre linéaire à matrices mal conditionnées et à données inexactes sont traités de diverses manières par Tikhonov (XVI ; 17], Faddeeva [XVI ; 19], M. M. Lavrentiev [II], Voévodine [VIII ; 3].

4.7. Méthodes directes de résolution des équations aux différences finies

De nombreux ouvrages parus au cours de la dernière décennie traitent des méthodes directes pour les systèmes d'équations aux différences finies. Il s'agit en premier lieu de la méthode de Fourier qui n'a été utilisée auparavant que dans des cas fort rares. Elle le cède en effet aux autres procédés quant au nombre d'opérations arithmétiques donnant la solution du problème. Le fait est qu'on devait calculer le système de fonctions propres, les coefficients et la somme de la série de Fourier.

La méthode de réduction cyclique constitue un outil très puissant en matière des équations aux différences finies (on trouve sa description assez complète dans Buzbee, Golub, Nielson [XII]). La réduction cyclique est en fait une variante originale de la méthode d'élimination de Gauss et un cas particulier de la factorisation.

4.7.1. Transformation de Fourier rapide

L'idée relative à FFT (Fast Fourier Transform) ne date pas d'hier, mais c'est Cooley et Tukey [XIII] qui ont mis la méthode à la mode en généralisant l'algorithme de Good et en réduisant sensiblement le nombre d'opérations nécessaires.

Soit une fonction $f(k)$ de l'argument discret, où le paramètre $k = 0, 1, 2, \ldots, N-1$. Représentons-la par la série finie (i.e. la somme) de Fourier

$$f(k) = \sum_{n=0}^{N-1} A(n)\, W^{kn}, \qquad A(n) = \frac{1}{N} \sum_{k=0}^{N-1} f(k)\, W^{-kn}. \tag{7.1}$$

On a introduit, comme dans Cooley et Tukey, la notation suivante pour la racine principale N-ième de l'unité:

$$W = e^{i2\pi/N}.$$

Nous appellerons opération deux opérations consécutives en arithmétique complexe, à savoir l'addition et la multiplication, auquel cas (7.1) implique que, connaissant $A(n)$ et W^{kn}, la recherche de $f(k)$ exige N^2 opérations.

D'après Cooley et Tukey, si N n'est pas un nombre premier, on réduit considérablement le nombre d'opérations en mettant (7.1) sous forme de somme multiple.

Prenons, en effet, le cas $N = N_1 \cdot N_2$, avec N_1 et N_2 des entiers naturels. Ecrivons k et n comme

$$k = k_1 N_2 + k_2; \quad k_1 = 0, 1, \ldots, N_1 - 1; \quad k_2 = 0, 1, \ldots, N_2 - 1; \tag{7.2}$$

$$n = n_1 + n_2 N_1; \quad n_1 = 0, 1, \ldots, N_1 - 1; \quad n_2 = 0, 1, \ldots, N_2 - 1.$$

Puisque

$$W^{k_1 n_2 N_1 N_2} = (W^N)^{k_1 n_2} = 1,$$

on a

$$W^{k n_2 N_1} = W^{k_2 n_2 N_1}$$

et

$$f(k) = f(k_1, k_2) = \sum_{n_1=0}^{N_1-1} \left[\sum_{n_2=0}^{N_2-1} A(n_1, n_2) W^{k_2 n_2 N_1} \right] W^{k n_1}. \quad (7.3)$$

Le problème de calculer la somme (7.1) se ramène donc à chercher la somme double (7.3) ou, ce qui revient au même, les sommes successives

$$A_1(n_1, k_2) = \sum_{n_2=0}^{N_2-1} A(n_1, n_2) W^{k_2 n_2 N_1}, \quad (7.4)$$

$$f(k_1, k_2) = \sum_{n_1=0}^{N_1-1} A_1(n_1, k_2) W^{k n_1}. \quad (7.5)$$

Il résulte cependant de (7.2) et (7.4) que le vecteur A_1 est N-dimensionnel et qu'on l'obtient par $N \times N_2$ opérations. Connaissant A_1, on trouve $f(k)$ à l'aide de (7.5) au bout de $N \times N_1$ opérations, i.e. elles sont au total $N \times (N_1 + N_2)$. Le nombre d'opérations diminue avec l'augmentation de N.

Si N_1 est premier et N_2 un nombre composé, on transforme la somme (7.4) par Cooley et Tukey (n_1 étant un paramètre) et on réduit le nombre d'opérations en factorisant N_2. Si $N = N_1 \times N_2 \times \ldots \times N_m$, ce nombre passe en général de N^2 à $N \times (N_1 + \ldots + N_m)$, la diminution la plus importante correspondant à $N_i = 2$, 3 ou 4. Avec $N = 256 = 2^8$, le nombre d'opérations diminue par exemple de $256/(8 \times 2) = 16$ fois et il devient $243/(5 \times 3) = 16{,}2$ fois moindre pour $N = 243 = 3^5$. Le cas le plus utile en programmation est celui de $N_i = 2$, $i = 1, 2, \ldots, m$, encore que $N_i = 4$, 8 soient à l'origine d'autres variantes économiques.

Soit donc $N = 2^m$. On pose $N_1 = 2$, $N_2 = 2^{m-1}$, on obtient des sommes du type (7.4) et (7.5) et on poursuit le processus. On a

$$k = \overline{k_{m-1} k_{m-2} \ldots k_1 k_0} \equiv k_{m-1} 2^{m-1} + k_{m-2} 2^{m-2} + \ldots + k_1 2 + k_0,$$

$$n = \overline{n_{m-1} n_{m-2} \ldots n_1 n_0} \equiv n_{m-1} 2^{m-1} + n_{m-2} 2^{m-2} + \ldots + n_1 2 + n_0,$$

où k_i et n_i valent 0 ou 1. Alors

$$f(k_{m-1}, \ldots, k_0) = \sum_{n_0=0}^{1} \left\{ \sum_{n_1=0}^{1} \left[\ldots \sum_{n_{m-1}=0}^{1} (A(n_{m-1}, \ldots, n_0) \times \right. \right.$$

$$\left. \left. \times W^{k n_{m-1} 2^{m-1}}) \ldots W^{k n_1 2} \right] W^{k n_0} \right\}. \quad (7.6)$$

Du moment que

$$W^{k n_{m-1} 2^{m-1}} = W^{k_0 n_{m-1} 2^{m-1}},$$

$$W^{k n_{m-2} 2^{m-2}} = W^{\overline{k_1 k_0} n_{m-2} 2^{m-2}},$$

et ainsi de suite, on ramène la recherche de la somme multiple (7.6) à celle d'une séquence de m sommes

$$A_1(k_0, n_{m-2}, \ldots, n_0) = \sum_{n_{m-1}=0}^{1} A(n_{m-1}, \ldots, n_0) W^{k_0 n_{m-1} 2^{m-1}},$$

$$A_2(k_1, k_0, n_{m-3}, \ldots, n_0) = \sum_{n_{m-2}=0}^{1} A_1(k_0, n_{m-2}, \ldots, n_0) \times \times W^{\overline{k_1 k_0} n_{m-2} 2^{m-2}}, \quad (7.7)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$A_m(k_{m-1}, \ldots, k_0) = \sum_{n_0=0}^{1} A_{m-1}(k_{m-2}, \ldots, k_0, n_0) W^{\overline{k_{m-1} \ldots k_0} n_0},$$

$$f(k) = A_m(k_{m-1}, \ldots, k_0).$$

Il y a lieu de noter que la transformation de Fourier rapide rend de grands services dans l'analyse statistique par corrélation des données pour les variables aléatoires $f(k)$, $k = 0, 1, \ldots, N-1$.

Abordons le problème de Dirichlet pour l'équation

$$-\Delta\varphi + \mu\varphi = f \text{ dans } D, \qquad \varphi = 0 \text{ sur } \partial D. \quad (7.8)$$

μ est une constante donnée et f une fonction suffisamment régulière définie dans $D = \{0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1\}$.

On associe à (7.8) le problème discrétisé

$$\frac{4\varphi_{k,l} - \varphi_{k-1,l} - \varphi_{k+1,l} - \varphi_{k,l-1} - \varphi_{k,l+1}}{h^2} + \mu\varphi_{k,l} = f_{k,l} \quad \text{dans} \quad D_h,$$

$$\varphi_{k,l} = 0 \quad \text{sur} \quad \partial D_h, \quad (7.9)$$

$$0 \leqslant k \leqslant \frac{1}{h} = N, \qquad 0 \leqslant l \leqslant \frac{1}{h} = N.$$

Si $\mu \geqslant 0$, les deux problèmes admettent chacun une solution et une seule. Si $\mu < 0$, il y a existence avec des contraintes supplémentaires sur μ et f.

Supposons qu'il y a existence et unicité et introduisons les notations

$$\varphi_l = \begin{Vmatrix} \varphi_{1,l} \\ \vdots \\ \varphi_{N-1\,l} \end{Vmatrix}, \quad f_l = \begin{Vmatrix} f_{1,l} \\ \vdots \\ f_{N-1,\,l} \end{Vmatrix}, \quad l = 1, \ldots, N-1,$$

$$A = \begin{Vmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{Vmatrix},$$

A étant une matrice d'ordre $N-1$. Désignons par E la matrice unité de même ordre.

Le problème (7.9) se récrit

$$\begin{aligned} B\varphi_1 - \varphi_2 &= h^2 f_1, \\ -\varphi_{l-1} + B\varphi_l - \varphi_{l+1} &= h^2 f_l, \quad l = 2, \ldots, N-2, \\ -\varphi_{N-2} + B\varphi_{N-1} &= h^2 f_{N-1}, \end{aligned} \tag{7.10}$$

avec $B = A + (2 + \mu h^2) E$.

On note que B et A possèdent une même base de vecteurs propres et que le problème de trouver toutes les valeurs propres

$$Au^{(m)} = \lambda_m (A)\, u^{(m)}$$

a pour solution

$$\lambda_m (A) = 2\left(1 - \cos\frac{m\pi}{N}\right), \qquad u_k^{(m)} = \sqrt{\frac{2}{N}} \sin\frac{m\pi k}{N},$$

avec $u_k^{(m)}$ la composante k-ième du vecteur propre $u^{(m)}$, $k = 1, 2, \ldots$ $\ldots, N-1, m = 1, 2, \ldots, N-1$. Le facteur $\sqrt{\frac{2}{N}}$ est introduit moyennant la condition de normalisation

$$\| u^{(m)} \|^2 = \sum_{k=1}^{N-1} (u_k^{(m)})^2 = 1.$$

Les vecteurs $u^{(m)}$ forment une base orthonormée dans l'espace de dimension $N-1$, si bien qu'on représente les vecteurs φ_l et f_l, $l = 1, \ldots, N-1$, par

$$\begin{aligned} \varphi_l &= \sum_{m=1}^{N-1} \Phi_{m,l}\, u^{(m)}, \\ f_l &= \sum_{m=1}^{N-1} F_{m,l}\, u^{(m)}. \end{aligned} \tag{7.11}$$

Portons dans le système (7.10) et multiplions membre à membre par $u^{(m)}$, il vient, pour tout m fixe, le système d'équations à matrice tridiagonale

$$\begin{aligned} \lambda_m \Phi_{m,1} - \Phi_{m,2} &= F_{m,1}, \\ -\Phi_{m,l-1} + \lambda_m \Phi_{m,l} - \Phi_{m,l+1} &= F_{m,l}, \quad l = 2, \ldots, N-2, \\ -\Phi_{m,N-2} + \lambda_m \Phi_{m,N-1} &= F_{m,N-1}. \end{aligned} \tag{7.12}$$

Ici $\lambda_m = \lambda_m (B) = \lambda_m (A) + 2 + \mu h^2$.

La résolution de (7.10) exige donc qu'on calcule les coefficients de Fourier de $N-1$ vecteurs f_l, la solution de $N-1$ systèmes à matrice tridiagonale de la forme (7.12) qui définissent les coefficients de Fourier des φ_l, $l = 1, \ldots, N-1$, et φ_l à l'aide de (7.11). Le développement de Fourier peut s'effectuer par FFT. A cette

fin, les formules donnant les coefficients de Fourier $F_{m,l}$ du vecteur f_l s'écrivent par exemple

$$F_{m,l} = \sqrt{\frac{2}{N}} \sum_{n=1}^{N-1} f_{n,l} \sin \frac{m\pi n}{N} = \sqrt{\frac{2}{N}} \sum_{n=0}^{2N-1} f_{n,l} \sin \frac{2m\pi n}{2N},$$

où $f_{0,l} = f_{N,l} = \ldots = f_{2N-1,l} = 0$. Notons $\tilde{w}$ la valeur de la racine principale M-ième ($M = 2N$) de l'unité. On a

$$F_{m,l} = \sqrt{\frac{2}{N}} \operatorname{Im} \left(\sum_{n=0}^{M-1} f_{n,l} \tilde{w}^{nm} \right), \qquad m = 1, \ldots, N-1,$$

et les sommes s'obtiennent directement par l'algorithme de Cooley et Tukey. On procède de même en ce qui concerne φ_l.

L'algorithme décrit de résolution directe de l'équation de Helmholtz avec la condition de Dirichlet joue dans la condition de Neumann ainsi que pour la condition « $\varphi(x, y)$ est périodique au contour du carré » et l'équation de la forme

$$a(y) \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial}{\partial y} \left(b(y) \frac{\partial \varphi}{\partial y} \right) - \mu(y) \varphi = f(x, y).$$

4.7.2. Méthode de réduction cyclique

Prenons une fois de plus le système (7.10) et supposons que $N = 2^{k+1}$. On rappelle que ce cas est le plus favorable pour calculer les séries finies de Fourier par l'algorithme de Cooley et Tukey. Or, ce système se prête pour $N = 2^{k+1}$ à une méthode directe comparable avec ledit algorithme pour ce qui est du nombre d'opérations arithmétiques. Nous voulons parler de la réduction cyclique dont la mise en œuvre s'effectue — et ceci constitue un avantage incontestable de la méthode, — sans qu'on connaisse les valeurs et vecteurs propres de la matrice B.

L'idée de la réduction cyclique est d'obtenir du système (7.10), N pair,

$$-\varphi_{l-1} + B\varphi_l - \varphi_{l+1} = h^2 f_l, \qquad l = 1, \ldots, N-1,$$

un système analogue où l'indice l de φ_l ne prend que les valeurs paires.

Ecrivons trois équations matricielles de (7.10):

$$-\varphi_{l-2} + B\varphi_{l-1} - \varphi_l = h^2 f_{l-1},$$
$$-\varphi_{l-1} + B\varphi_l - \varphi_{l+1} = h^2 f_l,$$
$$-\varphi_l + B\varphi_{l+1} - \varphi_{l+2} = h^2 f_{l+1},$$

l pair. Faisons la somme après avoir multiplié la deuxième équation terme à terme par B :

$$-\varphi_{l-2}+B^{(1)}\varphi_l-\varphi_{l+2}=f_l^{(1)}, \qquad l=2,\ 4,\ \ldots,\ N-2, \quad (7.13)$$

où $B^{(1)} = B^2 - 2E$, $f_l^{(1)} = f_{l-1} + Bf_l + f_{l+1}$, $l = 2,\ 4,\ \ldots, N-2$. On pose, pour simplifier les notations : $\varphi_0 = \varphi_N = f_0 = f_N = 0$.

On associe à chaque l impair pour φ_l (l pair) connus le système

$$B\varphi_l = h^2 f_l + \varphi_{l-1} + \varphi_{l+1}, \qquad l = 1,\ 3,\ \ldots,\ N-1. \quad (7.14)$$

Ainsi, nous avons *réduit* l'ordre du système d'équations matricielles. Le système (7.13) est également résoluble par le développement de Fourier parce que $B^{(1)}$ possède une base de vecteurs propres en commun avec B et A (voir p.ex. Hockney [XIII]).

Le système (7.13) comprend $\frac{N}{2} - 1$ équations matricielles. Comme $N = 2^{k+1}$, on procède comme plus haut et on aboutit à $\frac{N}{4} - 1$ équations de la forme

$$-\varphi_{4(l-1)}+B^{(2)}\varphi_{4l}-\varphi_{4(l+1)}=h^2 f_{4l}^{(2)}, \qquad l=1,\ 2,\ \ldots,\ N/4-1, \quad (7.15)$$

où

$$B^{(2)} = (B^{(1)})^2 - 2E,$$
$$f_{4l}^{(2)} = f_{4l-2}^{(1)} + f_{4l+2}^{(1)} + B^{(1)} f_{4l}^{(1)}.$$

On fait correspondre à chaque l pair non multiple de 4 pour φ_l connus avec l multiple de 4 le système

$$B^{(1)}\varphi_l = h^2 f^{(1)} + \varphi_{l-2} + \varphi_{l+2}, \qquad l=2,\ 6,\ 10,\ \ldots,\ N-2. \quad (7.16)$$

Après avoir répété k fois cette réduction cyclique, on a enfin

$$B^{(k)}\varphi_{2^k} = h^2 f_{2^k}^{(k)}, \quad (7.17)$$

et les autres inconnues s'obtiennent en résolvant les systèmes successifs

$$B^{(r)}\varphi_l = h^2 f_l^{(r)} + \varphi_{l-2^r} + \varphi_{l+2^r}, \quad (7.18)$$
$$l = (2i+1)\,2^r, \qquad i = 0,\ 1,\ \ldots,\ 2^{k-r}-1,$$
$$r = k-1,\ k-2,\ \ldots,\ 1,\ 0.$$

Les matrices $B^{(r)}$ vérifient les relations

$$B^{(0)} = B, \quad B^{(r+1)} = (B^{(r)})^2 - 2E, \qquad r = 0,\ 1,\ \ldots,\ k-1, \quad (7.19)$$

et les vecteurs $f_l^{(r)}$ sont définis par les formules

$$\begin{aligned} f_l^{(0)} &= f_l, \quad l = 1, \ldots, N-1; \\ f_l^{(r+1)} &= f_{l-2^r}^{(r)} + f_{l+2^r}^{(r)} + B^{(r)} f_l^{(r)}, \\ l &= j \cdot 2^{r+1}, \quad j = 1, 2, \ldots, 2^{k-r} - 1, \\ r &= 0, 1, \ldots, k-1. \end{aligned} \tag{7.20}$$

La matrice $B^{(r)}$ peut être factorisée en 2^r matrices tridiagonales. Considérons en effet la suite de polynômes

$$\begin{aligned} P_1(b) &= b, \\ P_{2^{r+1}}(b) &= (P_{2^r}(b))^2 - 2, \qquad r = 0, 1, \ldots \end{aligned}$$

Avec $b = 2\cos\varphi$, on calcule aisément

$$P_{2^r}(b) = 2\cos 2^r\varphi.$$

Par conséquent, $b_i = 2\cos\frac{(2i-1)\pi}{2^{r+1}}$, $i = 1, 2, \ldots, 2^r$, sont racines de $P_{2^r}(b)$ et la décomposition cherchée est de la forme

$$B^{(r)} = \prod_{i=1}^{2^r} \left(B - 2\cos\frac{(2i-1)\pi}{2^{r+1}} E\right). \tag{7.21}$$

Ainsi, on n'a pas à expliciter $B^{(r)}$ et la recherche de $x = (B^{(r)})^{-1} f$ sans calculer la matrice inverse ne demande que 2^r balayages des matrices tridiagonales (voir point 4.7.3).

On note que les composantes des vecteurs $f_l^{(r)}$ de (7.20) augmentent catastrophiquement avec les valeurs propres de $B^{(r)}$, d'où l'instabilité numérique. Un algorithme numériquement stable a été proposé en 1969 par Buneman (voir Buzbee, Golub, Nielson [XII]).

Dans la variante Buneman les vecteurs $f_l^{(r)}$ de (7.20) s'écrivent

$$\begin{aligned} h^2 f_l^{(r)} &= q_l^{(r)} - B^{(r)} p_l^{(r)}, \\ l &= i \cdot 2^r, \qquad i = 1, \ldots, 2^{k+1-r} - 1. \end{aligned} \tag{7.22}$$

Portons dans (7.20), il vient les formules suivantes pour calculer de proche en proche $p_l^{(r)}$ et $q_l^{(r)}$:

$$\begin{aligned} p_l^{(0)} &= 0, \quad q_l^{(0)} = h^2 f_l^{(0)}, \qquad l = 1, \ldots, 2^{k+1} - 1; \\ p_l^{(r+1)} &= p_l^{(r)} + (B^{(r)})^{-1} \left(p_{l-2^r}^{(r)} + p_{l+2^r}^{(r)} - q_l^{(r)}\right); \\ q_l^{(r+1)} &= q_{l-2^r}^{(r)} + q_{l+2^r}^{(r)} - 2p_l^{(r+1)}; \\ l &= i \cdot 2^{r+1}, \qquad i = 1, \ldots, 2^{k-r} - 1. \end{aligned} \tag{7.23}$$

Compte tenu de (7.22), on a, au lieu des systèmes matriciels (7.17) à (7.18),

$$B^{(r)}(\varphi_l + p_l^{(r)}) = q_l^{(r)} + \varphi_{l-2^r} + \varphi_{l+2^r},$$
$$l = (2i+1)\,2^r, \quad i = 0,\ 1,\ \ldots,\ 2^{k-r}-1, \qquad (7.24)$$
$$r = k,\ k-1,\ \ldots,\ 1,\ 0.$$

La réduction cyclique s'applique non seulement à l'équation de Helmholtz avec la condition de Dirichlet, mais aussi à la condition de Neumann et à celle « $\varphi(x, y)$ périodique au bord du carré » pour des équations plus complexes.

4.7.3. Factorisation des équations aux différences

On assiste depuis plusieurs années à la floraison de la littérature mathématique consacrée aux méthodes directes de résolution des équations aux différences finies. Parmi ces méthodes, la factorisation a fait ses preuves en tant que technique applicable avec succès à une vaste classe d'équations à une dimension. On en trouve les descriptions détaillées dans Vladimirov [XII; 4], Guelfand et Lokoutsievski [XII; 5], Godounov [XII; 7], Sauliev [III; 15], Ince [XII], Richtmyer [III], Marchouk et Lébédev [XVII; 12].

Il nous paraît inutile d'établir la priorité de tels auteurs au détriment de tels autres car il s'agit en fait du renouveau de bonnes vieilles idées à l'heure de l'analyse numérique moderne et du matériel électronique qui lui sert de support.

Soit le problème de diffusion simple

$$p\frac{d^2\varphi}{dx^2} - q\varphi = -f(x),$$
$$\frac{d\varphi}{dx} = 0 \quad \text{pour} \quad x = 0,$$
$$\varphi = 0 \quad \text{pour} \quad x = 1,$$

p et q étant des constantes positives et f une fonction continue donnée de l'argument x. Illustrons sur ce problème la méthode de factorisation.

On demande la solution continue dans $0 \leqslant x \leqslant 1$.

Considérons l'opérateur

$$L = p\frac{d^2}{dx^2} - q$$

et factorisons-le :

$$L = p\left(\frac{d}{dx} + \beta\right)\left(\frac{d}{dx} - \alpha\right),$$

avec α et β deux fonctions à définir moyennant la condition

$$p\left(\frac{d}{dx} + \beta\right)\left(\frac{d}{dx} - \alpha\right) \equiv p\frac{d^2}{dx^2} - q.$$

On vérifie cette identité sous la condition suffisante

$$\alpha = \beta,$$
$$\frac{d\beta}{dx} + \beta^2 = \frac{q}{p}.$$

Introduisons une nouvelle fonction z définie par la formule

$$z = -\frac{d\varphi}{dx} + \beta\varphi.$$

L'équation de diffusion se ramène au système efficace d'équations du premier ordre

$$\frac{d\beta}{dx} + \beta^2 = \frac{q}{p},$$
$$\frac{dz}{dx} + \beta z = \frac{f}{p},$$
$$\frac{d\varphi}{dx} - \beta\varphi = -z,$$

dit système factorisé.

On adjoint à ce système les conditions aux limites

$$\beta = 0 \quad \text{pour} \quad x = 0,$$
$$z = 0 \quad \text{pour} \quad x = 0,$$
$$\varphi = 0 \quad \text{pour} \quad x = 1$$

et on vérifie de suite qu'avec ce choix des données « initiales » du problème de Cauchy pour notre système, on est dans les conditions aux limites $\frac{d\varphi}{dx} = 0$ pour $x = 0$, $\varphi = 0$ pour $x = 1$ du problème proposé. Ainsi, le problème aux limites relatif à l'équation de diffusion fait place au problème de Cauchy pour trois équations différentielles ordinaires du premier ordre qu'on aborde l'une après l'autre. Il y a stabilité numérique si l'on calcule β et z dans le sens des x croissants et φ dans le sens contraire. Cette réduction d'un problème aux limites pour une équation du second ordre au problème de Cauchy relatif à des équations d'ordre 1 constitue la *méthode de factorisation*.

Le procédé décrit se généralise aisément au cas de problèmes de diffusion plus complexes à p, q, f discontinues par morceaux et avec diverses conditions aux limites.

Passons aux équations aux différences. L'algorithme de factorisation (appelé communément balayage) procède comme suit. On écrit l'équation aux différences de départ sous la forme

$$-a_i\varphi_{i-1} + p_i\varphi_i - c_i\varphi_{i+1} = f_i, \quad i = 1, 2, \ldots, n, \quad (7.25)$$

et on suppose avoir pris en considération les conditions aux limites pour φ cherchée, si bien que *)

$$a_1 = 0, \quad c_n = 0, \quad p_i \geqslant a_i + c_i. \tag{7.26}$$

L'équation (7.25) s'écrit sous forme matricielle

$$A\Phi = F, \tag{7.27}$$

où $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_n)^T$, $F = (f_1, f_2, \ldots, f_n)^T$. On récrit (7.27):

$$KS_1S_2\Phi = F, \tag{7.28}$$

avec

$$S_1 = \begin{Vmatrix} 1 & 0 & \ldots & 0 & 0 & 0 \\ -\alpha_2 & 1 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -\alpha_{n-1} & 1 & 0 \\ 0 & 0 & \ldots & 0 & -\alpha_n & 0 \end{Vmatrix},$$

$$S_2 = \begin{Vmatrix} 1 & -\xi_1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -\xi_2 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & -\xi_{n-1} \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{Vmatrix},$$

et K est une matrice diagonale qui joue le rôle de coefficient.

On introduit la notation $S_2\Phi = Z$, $K^{-1} = \Gamma$ et on obtient le système

$$S_1Z = \Gamma F, \quad S_2\Phi = Z. \tag{7.29}$$

Passons aux composantes de Z et Φ:

$$-\alpha_i z_{i-1} + z_i = \gamma_i f_i, \tag{7.30}$$

$$\varphi_i - \xi_i \varphi_{i+1} = z_i. \tag{7.31}$$

Pour trouver les coefficients α_i, ξ_i, γ_i, on porte dans (7.30) z_{i-1} et z_i de (7.31):

$$-\alpha_i\varphi_{i-1} + (1 + \alpha_i\xi_{i-1})\,\varphi_i - \xi_i\varphi_{i+1} = \gamma_i f_i.$$

On a par identification avec (7.25)

$$\begin{aligned} -\alpha_i &= \gamma_i a_i, \quad \xi_i = \gamma_i c_i, \\ \gamma_i &= (p_i - a_i\gamma_{i-1}c_{i-1})^{-1}. \end{aligned} \tag{7.32}$$

*) Les conditions aux limites discrétisées pour φ sont assimilables à des équations indépendantes.

Finalement, (7.30) et (7.31) s'écrivent

$$\begin{aligned} z_i &= \gamma_i (a_i z_{i-1} + f_i), \\ \varphi_i &= \gamma_i c_i \varphi_{i+1} + z_i. \end{aligned} \tag{7.33}$$

Il est connu (Godounov, Riabenki [III; 33, 34] que la stabilité numérique des équations (7.9) à (7.10) a lieu sous les conditions suffisantes $a_k \leqslant 0$, $c_k \leqslant 0$, $b_k \leqslant 0$, $a_k + c_k \leqslant b_k$, avec $a_k + c_k < b_k$ au moins pour un k.

Les formules de résolution de l'équation aux différences (7.25) s'écrivent d'ordinaire

$$\begin{aligned} \beta_{i+1} &= (p_i - a_i\beta_i)^{-1} c_i, \\ z_{i+1} &= (p_i - a_i\beta_i)^{-1} (a_i z_i + f_i), \\ \varphi_i &= \beta_{i+1}\varphi_{i+1} + z_{i+1}, \\ & \qquad i = 1, \ldots, n, \end{aligned} \tag{7.34}$$

si

$$\beta_1 = z_1 = 0, \quad \varphi_n = z_{n+1}.$$

La méthode de factorisation pour un système linéaire d'équations différentielles ordinaires du premier ordre avec contraintes linéaires quelconques (en particulier, avec des conditions à plusieurs points et des conditions aux limites) a été développée par Fagué [XII; 12].

L'algorithme (7.33), (7.34) joue également si φ_i, f_i de (7.25) sont des vecteurs et les coefficients a_i, c_i, p_i des matrices (à condition de ne pas changer l'ordre des facteurs dans ces formules).

On trouve la solution par les formules de récurrence (7.33), (7.34) compte tenu de (7.26).

S'agissant d'une équation aux différences elliptique de dimension 2, la factorisation matricielle s'avère efficace seulement si l'une des variables est donnée sous forme discrète en peu de points.

Bouléev [XII; 3] a proposé une méthode simple et suffisamment efficace pour les équations aux différences elliptiques en dimension 2 et 3 qu'il a appelée *factorisation incomplète*. Voici son idée.

Soit l'équation aux différences à deux dimensions

$$-a_{ik}\varphi_{i-1,\,k} - c_{ik}\varphi_{i+1,\,k} - b_{ik}\varphi_{i,\,k-1} - d_{ik}\varphi_{i,\,k+1} + p_{ik}\varphi_{ik} = f_{ik}, \tag{7.35}$$

$$i = 1, 2, \ldots, m; \quad k = 1, 2, \ldots, n,$$

$$a_{1k} = c_{mk} = b_{i1} = d_{in} = 0, \qquad p_{ik} \geqslant a_{ik} + c_{ik} + b_{ik} + d_{ik} \tag{7.36}$$

qui s'écrit sous forme matricielle

$$A\Phi = F,$$

où $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_N)^T$, $F = (f_1, f_2, \ldots, f_N)^T$, $N = mn$. Ajoutons aux deux membres le vecteur $B\Phi$, il vient

$$(A + B)\,\Phi = F + B\Phi. \tag{7.37}$$

On choisit une B telle que $A + B$ soit représentable par le produit de S_1 et S_2 tridiagonales simples et K diagonale :

$$A + B = KS_1S_2.$$

Par analogie avec (7.27), on remplace (7.37) par le système

$$\begin{aligned} S_1Z &= K^{-1}(F + B\Phi), \\ S_2\Phi &= Z, \end{aligned} \tag{7.38}$$

abordé par les approximations successives.

On substitue à (7.35) le système d'équations de la forme

$$\begin{aligned} z_{ik} &= \alpha_{ik}z_{i-1,\,k} + \beta_{ik}z_{i,\,k-1} + \gamma_{ik}[f_{ik} + D_{ik}(\varphi) - q_{ik}\varphi_{ik}], \\ \varphi_{ik} &= \xi_{ik}\varphi_{i+1,\,k} + \eta_{ik}\varphi_{i,\,k+1} + z_{ik}. \end{aligned} \tag{7.39}$$

Les coefficients α_{ik}, β_{ik}, ξ_{ik}, η_{ik}, γ_{ik} sont définis par des formules analogues à (7.32), (7.33), à savoir

$$\alpha_{ik} = \gamma_{ik}a_{ik}, \quad \beta_{ik} = \gamma_{ik}b_{ik}, \quad \xi_{ik} = \gamma_{ik}c_{ik}, \quad \eta_{ik} = \gamma_{ik}d_{ik},$$

$$\gamma_{ik} = (p_{ik} - q_{ik} - a_{ik}c_{i-1,\,k}\gamma_{i-1,\,k} - b_{ik}d_{i,\,k-1}\gamma_{i,\,k-1})^{-1},$$

et l'opérateur à intégrer $D_{ik}(\varphi)$ par

$$D_{ik}(\varphi) = a_{ik}\eta_{i-1,\,k}\varphi_{i-1,\,k+1} + b_{ik}\xi_{i,\,k-1}\varphi_{i+1,\,k-1}.$$

On prend $q_{ik} = \theta_{ik}(a_{ik}\eta_{i-1,\,k} + b_{ik}\xi_{i,\,k-1})$, $0 \leqslant \theta < 1$.

On garantit la convergence de (7.39) pour $\theta > 0$ par adjonction de l'itération simple

$$\widetilde{\varphi}_{ik}^{(l)} = \frac{1}{\gamma_{ik}p_{ik}}(\alpha_{ik}\varphi_{i-1,\,k}^{(l)} + \beta_{ik}\varphi_{i,\,k-1}^{(l)} + \xi_{ik}\varphi_{i+1,\,k}^{(l)} + \eta_{ik}\varphi_{i,\,k+1}^{(l)}), \tag{7.40}$$

si bien que la $(l + 1)$-ième approximation de φ_{ik} s'obtient en portant dans l'équation pour $z_{ik}^{(l+1)}$ la l-ième approximation $\widetilde{\varphi}_{ik}$.

On a formulé pour le schéma (7.39), (7.40) des critères suffisants de convergence.

Un schéma de factorisation incomplète proche de (7.39) a été proposé par Oliphant [XII].

Dans la suite Bouléev a construit un autre schéma défini par les relations

$$\begin{aligned} z_{ik} &= \alpha_{ik}z_{i-1,\,k} + \gamma_{ik}[f_{ik} + D_{ik}(\varphi) - q_{ik}\varphi_{ik}], \\ \varphi_{ik} - \beta_{ik}\varphi_{i,\,k-1} - \delta_{ik}\varphi_{i,\,k+1} &= \xi_{ik}\varphi_{i+1,\,k} + z_{ik} \end{aligned} \tag{7.41}$$

avec les coefficients α_{ik}, ξ_{ik}, β_{ik}, δ_{ik}, γ_{ik} et l'opérateur $D_{ik}(\varphi)$ tels que

$$\begin{aligned} &\alpha_{ik} = \gamma_{ik}a_{ik}, \quad \beta_{ik} = \gamma_{ik}b_{ik}, \quad \xi_{ik} = \gamma_{ik}c_{ik}, \quad \delta_{ik} = \gamma_{ik}d_{ik}, \\ &\gamma_{ik} = (p_{ik} - q_{ik} - a_{ik}c_{i-1,\,k}\gamma_{i-1,\,k})^{-1}, \\ &D_{ik}(\varphi) \equiv a_{ik}(\beta_{i-1,\,k}\varphi_{i-1,\,k-1} + \delta_{i-1,\,k}\varphi_{i-1,\,k-1}). \end{aligned} \tag{7.42}$$

Le coefficient q_{ik} est pris égal à $\theta a_{ik}(\beta_{i-1,k} + \delta_{i-1,\,k})$, $0 \leqslant \theta \leqslant 1$.

Une question se pose : comment réduire le « poids » relatif de l'expression à intégrer dans les équations (7.41), i.e. comment obtenir un opérateur aux différences $A + B$ suffisamment simple qui se prête à la factorisation et est « proche » de A donné?

A cet effet, Bouléev a quelque peu généralisé le schéma (7.41).

On cherche donc un système d'équations équivalent à (7.35) :

$$z_{ik} = \alpha_{ik} z_{i-1,\,k} + \mu_{ik} z_{i-1,\,k-1} + \nu_{ik} z_{i-1,\,k+1} + \\ + \gamma_{ik} [f_{ik} + D_{ik}(\varphi) - s_{ik}\varphi_{ik}], \qquad (7.43)$$

$$\varphi_{ik} - \beta_{ik}\varphi_{i,\,k-1} - \delta_{ik}\varphi_{i,\,k+1} = \xi_{ik}\varphi_{i+1,\,k} + z_{ik},$$

avec α_{ik}, β_{ik}, δ_{ik}, μ_{ik}, ν_{ik}, γ_{ik}, ξ_{ik} à déterminer.

Utilisons $z_{i-1,\,k}$, $z_{i-1,\,k-1}$ et $z_{i-1,\,k+1}$ de (7.43), il vient

$$-(\alpha_{ik} - \mu_{ik}\delta_{i-1,\,k-1} - \nu_{ik}\beta_{i-1,\,k+1})\,\varphi_{i-1,\,k} - \xi_{ik}\varphi_{i+1,\,k} - \\ - (\beta_{ik} - \mu_{ik}\xi_{i-1,\,k-1})\,\varphi_{i,\,k-1} - (\delta_{ik} - \nu_{ik}\xi_{i-1,\,k+1})\,\varphi_{i,\,k+1} + \\ + (1 + \alpha_{ik}\xi_{i-1,\,k} + \gamma_{ik}s_{ik})\,\varphi_{ik} + (\alpha_{ik}\beta_{i-1,\,k} - \mu_{ik})\,\varphi_{i-1,\,k-1} + \\ + (\alpha_{ik}\delta_{i-1,\,k} - \nu_{ik})\,\varphi_{i-1,\,k+1} + \mu_{ik}\beta_{i-1,\,k-1}\varphi_{i-1,\,k-2} + \\ + \nu_{ik}\delta_{i-1,\,k+1}\varphi_{i-1,\,k+2} - \gamma_{ik}D_{ik}(\varphi) = \gamma_{ik}f_{ik}. \qquad (7.44)$$

Les équations (7.44), (7.35) donnent 5 relations entre les coefficients α_{ik}, β_{ik}, . . ., μ_{ik}, ν_{ik}, γ_{ik} et ceux de (7.35) et l'expression de $D_{ik}(\varphi)$:

$$\begin{aligned} \alpha_{ik} - \mu_{ik}\delta_{i-1,\,k-1} - \nu_{ik}\beta_{i-1,\,k+1} &= \gamma_{ik}a_{ik}, \\ \xi_{ik} &= \gamma_{ik}c_{ik}, \\ \beta_{ik} - \mu_{ik}\xi_{i-1,\,k-1} &= \gamma_{ik}b_{ik}, \\ \delta_{ik} - \nu_{ik}\xi_{i-1,\,k+1} &= \gamma_{ik}d_{ik}, \end{aligned} \qquad (7.45)$$

$$1 + \alpha_{ik}\xi_{i-1,\,k} + \gamma_{ik}s_{ik} = \gamma_{ik}p_{ik}, \qquad (7.46)$$

$$\gamma_{ik}D_{ik}(\varphi) = (\alpha_{ik}\beta_{i-1,\,k} - \mu_{ik})\,\varphi_{i-1,\,k-1} + (\alpha_{ik}\delta_{i-1,\,k} - \nu_{ik})\,\varphi_{i-1,\,k+1} + \\ + \mu_{ik}\beta_{i-1,\,k-1}\varphi_{i-1,\,k-2} + \nu_{ik}\delta_{i-1,\,k+1}\varphi_{i-1,\,k+2}. \qquad (7.47)$$

Comme il y a cinq relations et sept coefficients inconnus, on en prend deux de façon quelconque. Posons

$$\mu_{ik} = \alpha_{ik}\beta_{i-1,\,k}, \qquad \nu_{ik} = \alpha_{ik}\delta_{i-1,\,k}, \qquad (7.48)$$

auquel cas les termes principaux de $D_{ik}(\varphi)$ s'annulent et les relations (7.46)-(7.47) s'écrivent

$$\begin{aligned} \alpha_{ik} &= \gamma_{ik}e_{ik}, \quad \text{où} \quad e_{ik} = \frac{a_{ik}}{1 - \beta_{i-1,\,k}\delta_{i-1,\,k-1} - \delta_{i-1,\,k}\beta_{i-1,\,k+1}}, \\ \xi_{ik} &= \gamma_{ik}c_{ik}, \end{aligned} \qquad (7.49)$$

$$\beta_{ik} = \gamma_{ik} b_{ik} + \alpha_{ik} \beta_{i-1,\,k} \xi_{i-1,\,k-1},$$
$$\delta_{ik} = \gamma_{ik} d_{ik} + \alpha_{ik} \delta_{i-1,\,k} \xi_{i-1,\,k+1}, \qquad (7.50)$$
$$\gamma_{ik} = (p_{ik} - s_{ik} - e_{ik} \xi_{i-1,\,k})^{-1},$$
$$\gamma_{ik} D_{ik}(\varphi) = \alpha_{ik} (\beta_{i-1,\,k} \beta_{i-1,\,k-1} \varphi_{i-1,\,k-2} + \delta_{i-1,\,k} \delta_{i-1,\,k+2}). \qquad (7.51)$$

On prend $\gamma_{ik} s_{ik}$ proportionnel à la somme des coefficients dans l'opérateur $\gamma_{ik} D_{ik}(\varphi)$:

$$\gamma_{ik} s_{ik} = \theta \alpha_{ik} (\beta_{i-1,\,k} \beta_{i-1,\,k-1} + \delta_{i-1,\,k} \delta_{i-1,\,k+1}), \qquad (7.52)$$
$$0 \leqslant \theta \leqslant 1.$$

On calcule γ_{ik} par

$$\gamma_{ik} = [p_{ik} - e_{ik} (\xi_{i-1,\,k} + \theta \beta_{i-1,\,k} \beta_{i-1,\,k-1} + \theta \delta_{i-1,\,k} \delta_{i-1,\,k+1})]^{-1} \qquad (7.53)$$

compte tenu de (7.52) et (7.49). Les coefficients de (7.43) sont cherchés dans l'ordre suivant :

$$e_{ik},\ \gamma_{ik},\ \alpha_{ik},\ \xi_{ik},\ \beta_{ik},\ \delta_{ik}.$$

On peut dire que la formule (7.53) régit le balayage explicite du vecteur

$$\gamma_i = (\gamma_{i1},\ \gamma_{i2},\ \ldots,\ \gamma_{in}).$$

L'équation (7.43) se récrit en définitive

$$z_{ik} = \alpha_{ik} (z_{i-1,\,k} + \beta_{i-1,\,k} z_{i-1,\,k-1} + \delta_{i-1,\,k} z_{i-1,\,k+1}) +$$
$$+ \gamma_{ik} f_{ik} + \alpha_{ik} [\beta_{i-1,\,k} \beta_{i-1,\,k-1} (\varphi_{i-1,\,k-2} - \theta \varphi_{ik}) +$$
$$+ \delta_{i-1,\,k} \delta_{i-1,\,k+1} (\varphi_{i-1,\,k+2} - \theta \varphi_{ik})], \qquad (7.54)$$
$$\varphi_{ik} - \beta_{ik} \varphi_{i,\,k-1} - \delta_{ik} \varphi_{i,\,k+1} = \xi_{ik} \varphi_{i+1,\,k} + z_{ik}.$$

On résout (7.53)-(7.54) par les approximations successives. On effectue d'abord le balayage explicite pour $z_i = (z_{i1},\ z_{i2},\ \ldots,\ z_{in})$, puis le balayage implicite par remontée pour $\Phi_i = (\varphi_{i1},\ \varphi_{i2},\ \ldots,\ \varphi_{in})$.

On note que le schéma (7.54) aboutit à la solution exacte pour Φ_i à deux composantes.

La résolution de l'équation de Poisson dans un domaine carré avec conditions aux frontières quelconques montre qu'il y a convergence des itérations (7.54) avec θ de l'intervalle $0 \leqslant \theta \leqslant 1$, la valeur optimale étant $\theta = 0{,}8\,;\ 0{,}9$.

Le Tableau 2 donne la norme q de l'opérateur de transition du processus itératif (7.54) en fonction du paramètre θ pour le problème de Dirichlet

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial \varphi^2}{\partial y^2} = -2(2 - x^2 - y^2),$$

$\varphi = 0$ pour $x = \pm 1$, $\varphi = 0$ pour $y = \pm 1$, le pas étant $\Delta x = \Delta y = \frac{1}{12}$ (529 nœuds).

Tableau 2

θ	0	0,3	0,5	0,6	0,7	0,8	0,85	0,9	0,95	1,0
q	0,86	0,82	0,76	0,72	0,65	0,47	0,43	0,49	0,61	1,0

Selon ce tableau, une itération (7.54) équivaut pour $\theta = 0{,}7$, 0,8 et 0,9 respectivement à 44, 77 ou 72 itérations simples.

Ainsi, le schéma (7.54) s'avère, pour θ optimal, plus efficace dans le cas du problème de Dirichlet proposé que les méthodes des directions alternées les plus performantes avec τ constant.

La fig. 17 représente les valeurs de première approximation relative à $y = 0$ qu'on calcule, conformément à ce schéma, pour $\theta = 0{,}8$,

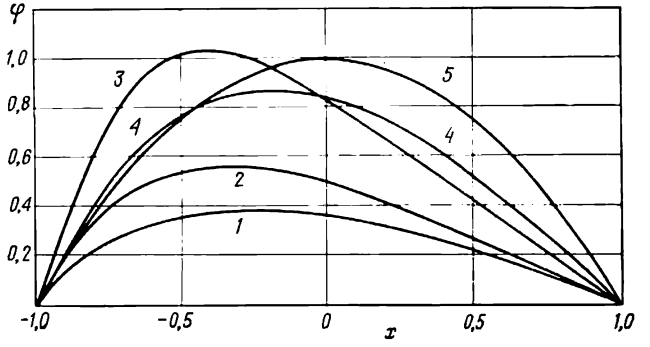


Fig. 17

$\theta = 0{,}9$ et $\theta = 1{,}0$ (la solution exacte est $\varphi = (1 - x^2)(1 - y^2)$). On voit que l'écart est minimal pour $\theta = 1{,}0$. Mais en seconde approximation la valeur optimale de θ est 0,9.

Les valeurs de sixième approximation correspondant à $\theta = 0{,}8$, $\theta = 0{,}85$ et $\theta = 0{,}9$ sont affectées d'erreurs moyennes en valeur absolue 0,013, 0,003 et 0,004 respectivement, i.e. on obtient pour $\theta = 0{,}85$ et $\theta = 0{,}9$ des solutions très proches de la solution stricte.

S'agissant du cas de Neumann, le schéma (7.54) le cède un peu aux meilleurs procédés utilisant les directions alternées.

Il est naturel que la première itération (7.54) diffère peu de la solution exacte si l'on pose θ égal à 1 ou sensiblement 1. C'est donc ces valeurs de θ qu'on prend dans un cas laborieux pour réaliser la (ou les) première itération (7.54), après quoi on passe à θ optimal.

Le fait de prendre, pour φ cherchée, des conditions aux limites arbitraires et un domaine bidimensionnel de forme quelconque n'alourdit nullement les schémas de factorisation incomplète. Mais

la condition aux limites discrétisée doit être de la forme (7.35) et telle que la somme de deux coefficients périphériques non nuls, quand ils sont positifs, soit au plus égale à p_{ik}. La factorisation incomplète l'emporte en ce sens sur les techniques des directions alternées qui deviennent d'emploi délicat si la frontière du domaine considéré ne coïncide pas avec les lignes de coordonnées du réseau.

Nous avons dit que les méthodes de factorisation incomplète sont très efficaces pour le problème de Dirichlet. Désireux d'en faire autant dans le cas de Neumann, Bouléev et Guinkine [XII; 3, 6] ont mis sur pied des schémas correspondants perfectionnés qui s'écrivent sous forme opératorielle (voir Bouléev [XII; 3])

$$\begin{aligned} z_{ik} &= \alpha_{ik} z_{i-1,\,k} + \gamma_{ik}\,[f_{ik} + D_{ik}(\varphi) - E_{ik}(\varphi)], \\ \varphi_{ik} - \beta_{ik}\varphi_{i,\,k-1} - \delta_{ik}\varphi_{i,\,k+1} &= \xi_{ik}\varphi_{i+1,\,k} + z_{ik}, \end{aligned} \tag{7.55}$$

où α_{ik}, β_{ik}, ξ_{ik}, δ_{ik}, γ_{ik} sont les coefficients à déterminer et toutes les φ de E proviennent du pochoir à 5 points.

Ce système équivaut à l'équation

$$\begin{aligned} (1 + \alpha_{ik}\xi_{i-1,\,k})\,\varphi_{ik} - \alpha_{ik}\varphi_{i-1,\,k} - \xi_{ik}\varphi_{i+1,\,k} - \beta_{ik}\varphi_{i,\,k-1} - \\ - \delta_{ik}\varphi_{i,\,k+1} + \alpha_{ik}\beta_{i-1,\,k}\varphi_{i-1,\,k-1} + \alpha_{ik}\delta_{i-1,\,k}\varphi_{i-1,\,k+1} - \\ - \gamma_{ik}D_{ik}(\varphi) + \gamma_{ik}E_{ik}(\varphi) = \gamma_{ik}f_{ik}. \end{aligned} \tag{7.56}$$

Les équations (7.56) et (7.35) entraînent

$$\gamma_{ik}D_{ik}(\varphi) = \alpha_{ik}\beta_{i-1,\,k}\varphi_{i-1,\,k-1} + \alpha_{ik}\delta_{i-1,\,k}\varphi_{i-1,\,k+1}. \tag{7.57}$$

On introduit $\gamma_{ik}E_{ik}$ défini par

$$\begin{aligned} \gamma_{ik}E_{ik}(\varphi) = \alpha_{ik}\beta_{i-1,\,k}(\varkappa\varphi_{i-1,\,k} + \eta\varphi_{i,\,k-1} + \sigma_i\omega\varphi_{i+1,\,k} - \theta_i\varphi_{ik}) + \\ + \alpha_{ik}\delta_{i-1,\,k}(\varkappa\varphi_{i-1,\,k} + \eta\varphi_{i,\,k+1} + \sigma_i\omega\varphi_{i+1,\,k} - \theta_i\varphi_{ik}), \end{aligned} \tag{7.58}$$

avec $\varkappa$, η, ω, θ_i des paramètres inconnus *),

$$\sigma_i = \begin{cases} 1 & \text{pour } i = 1, 2, \ldots, m-1, \\ 0 & \text{pour } i = m. \end{cases}$$

Les termes en φ_{ik} de (7.58) servent à « renforcer » la diagonale principale de la matrice $(A + D - E)$ qui se factorise, si bien que la combinaison linéaire $\varkappa\varphi_{i-1,\,k} + \eta\varphi_{i,\,k-1} + \sigma_i\omega\varphi_{i+1,\,k}$ compense la somme $\varphi_{i-1,\,k-1} + \theta_i\varphi_{ik}$ et la combinaison $\varkappa\varphi_{i-1,\,k} + \eta\varphi_{i,\,k+1} + \sigma_i\omega\varphi_{i+1,\,k}$ la somme $\varphi_{i-1,\,k+1} + \theta_i\varphi_{ik}$.

*) Le fait d'introduire le facteur $\sigma_m = 0$ risque d'altérer la formule générale de $E_{ik}(\varphi)$. On assimile donc les conditions aux limites à droite

$$-a_{m+1,\,k}\varphi_{m,\,k} + \varphi_{m+1,\,k} = f_{m+1,\,k} \tag{7.59}$$

à une colonne d'équations indépandantes, auquel cas on omet de calculer $D_{ik}(\varphi)$, $E_{ik}(\varphi)$ et z_{ik} de la colonne $i = m + 1$ et on cherche la solution de (7.55) à partir de la colonne m après qu'on y élimine $\varphi_{m+1,\,k}$ moyennant les conditions (7.59).

On assujettit les paramètres $\varkappa$, η, ω, θ_i à vérifier la condition

$$1 + \theta_i = \varkappa + \eta + \sigma_i\omega + \varepsilon, \tag{7.60}$$

ε étant petit positif ou nul.

Les équations (7.56) et (7.35) conduisent, compte tenu de (7.58), aux relations suivantes entre les coefficients de (7.55) et ceux de (7.35):

$$1 + \alpha_{ik}\xi_{i-1,\,k} - \theta_i\alpha_{ik}(\beta_{i-1,\,k} + \delta_{i-1,\,k}) = \gamma_{ik}p_{ik}, \tag{7.61}$$

$$\begin{aligned} \alpha_{ik} - \varkappa\alpha_{ik}(\beta_{i-1,\,k} + \delta_{i-1,\,k}) &= \gamma_{ik}a_{ik}, \\ \beta_{ik} - \eta\alpha_{ik}\beta_{i-1,\,k} &= \gamma_{ik}b_{ik}, \\ \delta_{ik} - \eta\alpha_{ik}\delta_{i-1,\,k} &= \gamma_{ik}d_{ik}, \\ \xi_{ik} - \sigma_i\omega\alpha_{ik}(\beta_{i-1,\,k} + \delta_{i-1,\,k}) &= \gamma_{ik}c_{ik} \end{aligned} \tag{7.62}$$

ou

$$\begin{aligned} \gamma_{ik} &= [p_{ik} + (1 - \varkappa\beta_{i-1,\,k} - \varkappa\delta_{i-1,\,k})^{-1}\, a_{ik}(\theta_i\beta_{i-1,\,k} + \theta_i\delta_{i-1,\,k} - \xi_{i-1,\,k})]^{-1}, \\ \alpha_{ik} &= [1 - \varkappa(\beta_{i-1,\,k} + \delta_{i-1,\,k})]^{-1}\gamma_{ik}a_{ik}, \\ \beta_{ik} &= \gamma_{ik}b_{ik} + \eta\alpha_{ik}\beta_{i-1,\,k}, \\ \delta_{ik} &= \gamma_{ik}d_{ik} + \eta\alpha_{ik}\delta_{i-1,\,k}, \\ \xi_{ik} &= \gamma_{ik}c_{ik} + \sigma_i\omega\alpha_{ik}(\beta_{i-1,\,k} + \delta_{i-1,\,k}). \end{aligned} \tag{7.63}$$

Le schéma itératif (7.55) s'écrit par suite de (7.57) et (7.58)

$$\begin{aligned} z_{ik}^{(l)} &= \alpha_{ik}z_{i-1,\,k}^{(l)} + \gamma_{ik}f_{ik} + \\ &+ \alpha_{ik}\beta_{i-1,\,k}(\varphi_{i-1,\,k-1}^{(l-1)} + \theta_i\varphi_{ik}^{(l-1)} - \varkappa\varphi_{i-1,\,k}^{(l-1)} - \eta\varphi_{i,\,k-1}^{(l-1)} - \sigma_i\omega\varphi_{i+1,\,k}^{(l-1)}) + \\ &+ \alpha_{ik}\delta_{i-1,\,k}(\varphi_{i-1,\,k+1}^{(l-1)} + \theta_i\varphi_{ik}^{(l-1)} - \varkappa\varphi_{i-1,\,k}^{(l-1)} - \eta\varphi_{i,\,k+1}^{(l-1)} - \sigma_i\omega\varphi_{i+1,\,k}^{(l-1)}), \end{aligned} \tag{7.64}$$

$$\varphi_{ik}^{(l)} - \beta_{ik}\varphi_{i,\,k-1}^{(l)} - \delta_{ik}\varphi_{i,\,k+1}^{(l)} = \xi_{ik}\varphi_{i+1,\,k}^{(l)} + z_{ik}^{(l)}. \tag{7.65}$$

On résout (7.65) par balayage à une dimension

$$\rho_{ik} = (1 - \beta_{ik}\delta_{i,\,k-1}\rho_{i,\,k-1})^{-1},$$

$$w_{ik} = \rho_{ik}(\beta_{ik}w_{i,\,k-1} + \xi_{ik}\varphi_{i+1,\,k}^{(l)} + z_{ik}^{(l)}), \tag{7.66}$$

$$\varphi_{ik} = \rho_{ik}\delta_{ik}\varphi_{i,\,k+1} + w_{ik}. \tag{7.67}$$

On prend naturellement (7.55) tel que la somme des coefficients de l'opérateur $D_{ik}(\varphi) - E_{ik}(\varphi)$ soit proche de 0 et celle de leurs modules soit le plus petite possible par rapport à a_{ik}, coefficient de l'équation de départ (7.35).

Nous dirons que l'opérateur $(A + B)_{ik}$ est voisin de A_{ik} au sens de l'approximation si la somme des coefficients de B_{ik} est proche de 0, et qu'il l'est pour la norme du résidu si la somme de leurs modules est petite par rapport aux valeurs du coefficient a_{ik} de A_{ik}.

La première propriété est garantie pour notre schéma par la relation (7.60).

La valeur optimale de θ doit être commensurable avec celle du coefficient précédant $\varphi_{i-1,\,k-1}$ et $\varphi_{i-1,\,k+1}$, i.e. elle doit être de l'ordre de l'unité. Si l'on se propose d'obtenir la proportionnalité approchée entre β_{ik}, δ_{ik} et ξ_{ik}, d'une part, et b_{ik}, d_{ik} et c_{ik}, de l'autre, on relie η et ω, selon (7.63), par l'égalité approchée

$$2\omega \approx \eta.$$

Etudions le schéma (7.64), (7.65) sous l'angle de la stabilité numérique dans l'espace.

On montre par récurrence que les coefficients β_{ik}, δ_{ik} et ξ_{ik} remplissent la condition

$$\beta_{ik} + \delta_{ik} + \xi_{ik} \leqslant 1. \tag{7.68}$$

En effet, la relation (7.60) donne lieu pour $i = 1, 2, \ldots, m-1$, par suite de (7.36) et (7.61), à

$$\begin{aligned} 1 + \alpha_{ik}\xi_{i-1,\,k} - \theta\alpha_{ik}(\beta_{i-1,\,k} + \delta_{i-1,\,k}) \geqslant \\ \geqslant \alpha_{ik} - \varkappa\alpha_{ik}(\beta_{i-1,\,k} + \delta_{i-1,\,k}) + \beta_{ik} - \varkappa\alpha_{ik}\beta_{i-1,\,k} + \\ + \delta_{ik} - \eta\alpha_{ik}\delta_{i-1,\,k} + \xi_{ik} + \omega\alpha_{ik}(\beta_{i-1,\,k} + \delta_{i-1,\,k}), \end{aligned} \tag{7.69}$$

ou

$$\begin{aligned} 1 - \xi_{ik} - \beta_{ik} - \delta_{i,\,k} \geqslant \alpha_{ik}[1 - \xi_{i-1,\,k} - \beta_{i-1,\,k} - \\ - \delta_{i-1,\,k} + (1 + \theta - \varkappa - \eta - \omega)(\beta_{i-1,\,k} + \theta_{i-1,\,k})] \end{aligned}$$

dont le second membre est positif par hypothèse, d'où la validité de (7.68).

Voyons la propagation des erreurs dans le schéma itératif (7.64), (7.65). Ecrivons la formule de récurrence pour $\frac{z_{ik}}{\xi_{ik}}$. On a pour $i = 1, 2, \ldots, m-1$:

$$\frac{z_{ik}}{\xi_{ik}} = \frac{\alpha_{ik}}{\xi_{ik}}\xi_{i-1,\,k}\frac{z_{i-1,\,k}}{\xi_{i-1,\,k}} + \frac{\gamma_{ik}}{\xi_{ik}}f_{ik}.$$

Compte tenu de (7.68) et (7.63),

$$\begin{aligned} \frac{\alpha_{ik}}{\xi_{ik}}\xi_{i-1,\,k} \leqslant \frac{\alpha_{ik}}{\xi_{ik}}(1 - \beta_{i-1,\,k} - \delta_{i-1,\,k}) = \\ = \frac{a_{ik}[1 - (\beta_{i-1,\,k} + \delta_{i-1,\,k})]}{c_{ik}[1 - \varkappa(\beta_{i-1,\,k} + \delta_{i-1,\,k})] + \omega a_{ik}(\beta_{ik} + \delta_{ik})} \end{aligned} \tag{7.70}$$

Si l'on pose $\varkappa \leqslant 1$ et si l'on note

$$\beta_{i-1,\,k} + \delta_{i-1,\,k} = q_{i-1,\,k},$$

on obtient

$$\frac{\alpha_{ik}}{\xi_{ik}}\,\xi_{i-1,\,k} \leqslant \frac{a_{ik}}{c_{ik}+\omega a_{ik}s_{i-1,\,k}}\,, \qquad s_{ik}=\frac{q_{ik}}{1-q_{ik}}\,. \tag{7.71}$$

On rend le second membre inférieur à 1 par un bon choix de ω *).

Ainsi, le schéma (7.64)-(7.65) est stable par rapport à la variable d'espace.

Les expériences numériques ont fourni les bonnes valeurs de $\varkappa$, η, ω et ε pour les problèmes de Dirichlet et de Neumann concernant l'équation de Poisson :

$$\varkappa = 0{,}5 \div 1, \quad \eta = 1, \quad \omega = 0 \div 0{,}5, \quad \varepsilon = 0 \qquad (\varkappa + \omega \approx 1). \tag{7.72}$$

*) La chose est importante pour le calcul des écoulements fluides turbulents où la viscosité et la conductibilité thermique varient de façon considérable.

CHAPITRE 5

MÉTHODES DE RÉSOLUTION DES PROBLÈMES NON STATIONNAIRES

Ce chapitre portera essentiellement sur le problème d'évolution de la physique mathématique

$$\frac{\partial \varphi}{\partial t} + A\varphi = f \text{ dans } D \times D_t, \qquad \varphi = g \text{ dans } D \text{ pour } t = 0,$$

où $A \geqslant 0$ et la solution φ présente la régularité nécessaire, ainsi que les fonctions f et g. On suppose que la solution vérifie sur ∂D certaines conditions aux limites *).

5.1. Schémas aux différences d'ordre d'approximation 2 à opérateurs dépendant du temps

Soit l'équation d'évolution

$$\frac{\partial \varphi}{\partial t} + A\varphi = 0 \text{ dans } D \times D_t, \qquad \varphi = g \text{ dans } D \text{ pour } t = 0. \tag{1.1}$$

Son analogue aux différences s'écrit

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\,\frac{\varphi^{j+1} + \varphi^j}{2} = 0, \qquad \varphi^0 = g. \tag{1.2}$$

On vérifie sans peine que sous la condition de régularité suffisante de la solution, le premier problème est approché par (1.2) à l'ordre 2 en τ. L'équation (1.2) connue sous le nom de schéma de Crank-Nicholson s'obtient en appliquant à tour de rôle les schémas explicite et implicite exacts à l'ordre 1 associés aux intervalles respectifs $t_j \leqslant t \leqslant t_{j+1/2}$ et $t_{j+1/2} \leqslant t \leqslant t_{j+1}$ (si A est un opérateur linéaire indépendant de t):

$$\begin{aligned} \frac{\varphi^{j+1/2} - \varphi^j}{\tau/2} + A\varphi^j &= 0, \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau/2} + A\varphi^{j+1} &= 0. \end{aligned} \tag{1.3}$$

On arrive au schéma de Crank-Nicholson par élimination des inconnues $\varphi^{j+1/2}$.

*) Sauf mention du contraire, l'opérateur A sera supposé discrétisé.

On suppose maintenant que A de (1.1) dépend du temps et est approché par un opérateur aux différences Λ. Il s'agit donc d'un problème d'algèbre linéaire et

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda^j\frac{\varphi^{j+1}+\varphi^j}{2}=0, \qquad \varphi^0=g, \tag{1.4}$$

$$(\Lambda^j\varphi,\ \varphi)\geqslant 0 \tag{1.5}$$

quelles que soient les fonctions de Φ.

Résolvons (1.4) par rapport à φ^{j+1}, il vient

$$\varphi^{j+1}=\left(E+\frac{\tau}{2}\Lambda^j\right)^{-1}\left(E-\frac{\tau}{2}\Lambda^j\right)\varphi^j \tag{1.6}$$

ou

$$\varphi^{j+1}=T^j\varphi^j, \tag{1.7}$$

T^j étant l'opérateur de transition

$$T^j=\left(E+\frac{\tau}{2}\Lambda^j\right)^{-1}\left(E-\frac{\tau}{2}\Lambda^j\right). \tag{1.8}$$

On démontre la stabilité numérique sans évaluer la norme de T^j.

Multiplions (1.4) scalairement par $\frac{1}{2}(\varphi^{j+1}+\varphi^j)$:

$$\frac{(\varphi^{j+1},\ \varphi^{j+1})-(\varphi^j,\ \varphi^j)}{2\tau}+\left(\Lambda^j\frac{\varphi^{j+1}+\varphi^j}{2},\ \frac{\varphi^{j+1}+\varphi^j}{2}\right)=0. \tag{1.9}$$

L'opérateur Λ^j est semi-défini positif par hypothèse (voir (1.5)), si bien que

$$\|\varphi^{j+1}\|\leqslant\|\varphi^j\|, \tag{1.10}$$

i.e. il y a stabilité.

L'évaluation en question constitue néanmoins un élément précieux de l'analyse des schémas aux différences. Elle est obtenue moyennant la relation (1.10) et la définition de la norme de l'opérateur (voir n° 1.1):

$$\|T^j\|\leqslant 1. \tag{1.11}$$

Ce résultat s'obtient d'ailleurs de façon directe par application du lemme de Kellogg (voir n° 1.1), de (1.8) et des inégalités $\Lambda^j\geqslant 0$, $\tau>0$. Si l'opérateur Λ^j est symétrique gauche, i.e.

$$(\Lambda^j\varphi,\ \varphi)=0,$$

on a, au lieu de (1.10),

$$\|\varphi^{j+1}\|=\|\varphi^j\|, \tag{1.12}$$

et on montre (comme on vient de lefaire) que

$$\|T^j\|=1. \tag{1.13}$$

Reprenons le schéma de Crank-Nicholson pour A dépendant de t.

On introduit l'opérateur H défini par

$$H\varphi \equiv \frac{\partial \varphi}{\partial t} = A\varphi \tag{1.14}$$

et l'opérateur H_τ défini par

$$(H_\tau \varphi)^j = \frac{(\varphi)^{j+1} - (\varphi)^j}{\tau} + \Lambda^j \frac{(\varphi)^{j+1} + (\varphi)^j}{2}, \tag{1.15}$$

$(\varphi)^j$ étant la projection de la fonction φ sur le réseau D_τ. On introduit de plus la norme

$$\| (H_\tau \varphi) \|_{C_\tau} = \max_{tj \in D_\tau} \| (H_\tau \varphi)^j \|, \tag{1.16}$$

où $\| \cdot \|$ est une norme sur l'espace contenant $(H_\tau\ \varphi_i)$. Proposons-nous d'évaluer la norme (1.16). On développe la solution de (1.1) en série de Taylor:

$$(\varphi)^{j+1} = (\varphi)^j + \tau\, (\varphi_t)^j + \frac{\tau^2}{2} (\varphi_{tt})^j + \ \ldots \tag{1.17}$$

Etant données les relations évidentes

$$\varphi_t = -A\varphi, \qquad \varphi_{tt} = A^2\varphi - A_t\varphi, \tag{1.18}$$

$A_t = \frac{\partial A}{\partial t}$, on transforme (1.17) en

$$(\varphi)^{j+1} = (\varphi)^j - \tau A^j (\varphi)^j + \frac{\tau^2}{2} [(A^j)^2 (\varphi)^j - A_t^j (\varphi)^j] - \ \ldots \tag{1.19}$$

Portons dans (1.16), il vient, compte tenu de (1.15),

$$\| (f - H_\tau \varphi) \|_{C_\tau} =$$

$$= \max_{t_j} \| \Lambda^j (\varphi)^j - A^j (\varphi)^j + \frac{\tau}{2} \{(A^j)^2 - A_t^j - \Lambda^j A^j\} (\varphi)^j + O(\tau^2) \|, \tag{1.20}$$

avec f^j le second membre de (1.1) qui est en l'occurrence égal à 0.

Si l'on prend l'opérateur d'approximation Λ^j

$$\Lambda^j = A^j = A\ (t_j), \tag{1.21}$$

alors (1.20) entraîne

$$\| (f - H_\tau \varphi) \|_{C_\tau} = \frac{\tau}{2} \max_{t_j \in D_\tau} \| A_t^j (\varphi)^j \| + O(\tau^2),$$

et l'approximation est d'ordre 1. S'agissant d'un cas particulier de A indépendant de t, l'opérateur (1.21) approche à l'ordre 2 en τ.

Soit maintenant

$$\Lambda^j = A^j + \frac{\tau}{2} A_t^j, \tag{1.22}$$

auquel cas

$$\| (f - H_\tau \varphi) \|_{C_\tau} = O(\tau^2).$$

On note que le schéma de Crank-Nicholson approche à l'ordre 2 en τ pour Λ^j tel que

$$\Lambda^j = A^{j+1/2} \tag{1.23}$$

ou

$$\Lambda^j = \frac{1}{2}(A^{j+1} + A^j). \tag{1.24}$$

On utilise dans les applications, notamment pour calculer les solutions des équations quasi linéaires, les approximations (1.22), (1.23) ou (1.24) qui donnent une précision d'ordre 2.

5.2. Equations d'évolution non homogènes

Dans le numéro précédent nous avons considéré des équations homogènes. Soit

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + A\varphi = f \quad \text{dans} \quad D \times D_t, \\ &\varphi = g \quad \text{dans} \quad D \quad \text{pour} \quad t = 0. \end{aligned} \tag{2.1}$$

Son approximation aux différences par le schéma de Crank-Nicholson sous les hypothèses du n° 4.1 est de la forme

$$\begin{aligned} &\frac{\varphi^{j+1} - \varphi^j}{\tau} + \Lambda^j \frac{\varphi^{j+1} + \varphi^j}{2} = f^j, \\ &\varphi^0 = g, \end{aligned} \tag{2.2}$$

où

$$f^j = f(t_{j+1/2}).$$

Il est immédiat de vérifier qu'elle approche (2.1) à l'ordre 2 en τ. La solution de (2.2) s'écrit

$$\varphi^{j+1} = T^j \varphi^j + \tau \left(E + \frac{\tau}{2} \Lambda^j \right)^{-1} f^j. \tag{2.3}$$

On a établi (voir n° 5.1) la majoration suivante pour l'équation homogène avec $\Lambda^j \geqslant 0$:

$$\| T^j \| \leqslant 1. \tag{2.4}$$

Il découle de (2.3) que

$$\| \varphi^{j+1} \| \leqslant \| T^j \| \| \varphi^j \| + \tau \left\| \left(E + \frac{\tau}{2} \Lambda^j \right)^{-1} \right\| \| f^j \|. \tag{2.5}$$

Abordons le problème de stabilité. Utilisons l'estimation (1.25) du Chapitre premier. Puisque $\tau > 0$ et

$$(\Lambda^j \varphi^j, \varphi^j) \geqslant 0, \tag{2.6}$$

on a

$$\left\| \left(E + \frac{\tau}{2} \Lambda^j \right)^{-1} \right\| \leqslant 1, \tag{2.7}$$

et l'inégalité (2.5) devient, par suite de (2.4) et (2.7),

$$\| \varphi^{j+1} \| \leqslant \| \varphi^j \| + \tau \| f^j \|. \tag{2.8}$$

Posons $\| \varphi^0 \| = \| g \|$ et $\| f \| = \max_j \| f^j \|$, il vient à l'aide de la relation de récurrence (2.8)

$$\| \varphi^j \| \leqslant \| g \| + j\tau \| f \|, \qquad j\tau \leqslant \text{const}, \tag{2.9}$$

i.e. le schéma aux différences est stable. La relation (2.9) constitue de plus une estimation à priori de la norme de la solution.

5.3. Méthodes de décomposition des problèmes non stationnaires

De nombreux problèmes volumineux de la physique mathématique se ramènent à une suite de problèmes plus simples qui se prêtent bien au calcul sur ordinateur. Cette réduction a lieu si l'opérateur semi-défini positif du problème initial se met sous forme de somme d'opérateurs semi-définis positifs élémentaires. Ces méthodes sont connues sous le nom de méthodes de décomposition. Les promoteurs en ont été, dans le cas non stationnaire, Douglas, Peaceman et Rachford (voir [XV]) qui ont été suivis par les mathématiciens soviétiques Bagrinovski et Godounov [XV; 2], Yanenko [III; 56] [XV; 13], Samarski [III; 12] [XV; 11], Diakonov [III; 6] [XV; 5], Sauliev [III; 15], Marchuk [XV].

Les méthodes de décomposition ont été primitivement formulées et légitimées pour des problèmes simples à opérateurs définis positifs commutatifs. On devait comprendre plus tard que dans ce cas toutes leurs variantes sont ou bien équivalentes (à part le schéma de mise en œuvre), ou bien voisines.

Depuis, la gamme de problèmes non triviaux justiciables de la décomposition s'est étendue tant et si bien que ces méthodes s'appliquent avec succès à des cas très complexes relevant de la physique mathématique. Nous commencerons par le cas le plus étudié où l'opérateur se décompose en deux opérateurs plus simples. Le lecteur est prié de ne pas oublier que c'est la méthode de splitting-up qui est, selon nous, la plus utile dans les applications.

Soit donc l'équation d'évolution

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + A\varphi = f \quad \text{dans} \quad D \times D_t, \\ &\varphi = g \,\text{dans}\, D \quad \text{pour} \quad t = 0, \end{aligned} \tag{3.1}$$

où l'opérateur $A \geqslant 0$ ne dépend pas du temps et admet la représentation

$$A = A_1 + A_2 \tag{3.2}$$

si

$$A_1 \geqslant 0, \qquad A_2 \geqslant 0. \tag{3.3}$$

Admettons que la solution du problème (3.1) est suffisamment régulière. On supposera de plus, si besoin est, que (3.1) est mis sous forme de différences finies, si bien que A, A_1 et A_2 sont des matrices.

5.3.1. Méthode de stabilisation

Considérons le schéma aux différences associé au problème (3.1)-(3.3) avec $f = 0$

$$\left(E + \frac{\tau}{2} A_1\right)\left(E + \frac{\tau}{2} A_2\right)\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^j = 0, \qquad \varphi^0 = g. \tag{3.4}$$

On montre aisément que sous l'hypothèse de solution suffisamment régulière, le problème (3.4) approche (3.1)-(3.3) à des quantités petites du second ordre près en τ. En effet, on ramène (3.4) par transformations algébriques à

$$\left(E + \frac{\tau^2}{4} A_1 A_2\right)\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\frac{\varphi^{j+1} + \varphi^j}{2} = 0, \qquad \varphi^0 = g. \tag{3.5}$$

On constate de suite qu'étant donnée la régularité suffisante de la solution, l'équation aux différences (3.5) est du même ordre d'approximation que le schéma de Crank-Nicholson

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\frac{\varphi^{j+1} + \varphi^j}{2} = 0, \qquad \varphi^0 = g, \tag{3.6}$$

car (3.6) approche à l'ordre 2 en τ.

Voyons si (3.4) est stable. On met cette équation sous la forme

$$\left(E + \frac{\tau}{2} A_1\right)\left(E + \frac{\tau}{2} A_2\right)\varphi^{j+1} =$$

$$= \left(E - \frac{\tau}{2} A_1\right)\left(E - \frac{\tau}{2} A_2\right)\varphi^j, \qquad \varphi^0 = g, \tag{3.7}$$

et on résout par rapport à φ^{j+1}:

$$\varphi^{j+1} = \left(E + \frac{\tau}{2} A_2\right)^{-1}\left(E + \frac{\tau}{2} A_1\right)^{-1}\left(E - \frac{\tau}{2} A_1\right)\left(E - \frac{\tau}{2} A_2\right)\varphi^j. \tag{3.8}$$

On passe de φ^j à ψ^j par la formule

$$\psi^j = \left(E + \frac{\tau}{2} A_2\right)\varphi^j, \tag{3.9}$$

et on obtient pour la nouvelle inconnue

$$\psi^{j+1} = T\psi^j, \tag{3.10}$$

T étant l'opérateur de transition

$$T = \left(E + \frac{\tau}{2} A_1\right)^{-1} \left(E - \frac{\tau}{2} A_1\right) \left(E - \frac{\tau}{2} A_2\right) \left(E + \frac{\tau}{2} A_2\right)^{-1}. \tag{3.11}$$

On trouve moyennant (3.10)

$$\|\psi^{j+1}\| \leqslant \|T\| \, \|\psi^j\|. \tag{3.12}$$

La norme de T est majorée par l'inégalité

$$\|T\| \leqslant \|T_1\| \, \|T_2\|,$$

où

$$T_\alpha = \left(E - \frac{\tau}{2} A_\alpha\right) \left(E + \frac{\tau}{2} A_\alpha\right)^{-1}, \qquad \alpha = 1, 2. \tag{3.13}$$

On a utilisé la propriété

$$\left(E + \frac{\tau}{2} A_\alpha\right)^{-1} \left(E - \frac{\tau}{2} A_\alpha\right) = \left(E - \frac{\tau}{2} A_\alpha\right) \left(E + \frac{\tau}{2} A_\alpha\right)^{-1}$$

qui découle de l'identité évidente

$$\left(E + \frac{\tau}{2} A_\alpha\right) \left(E + \frac{\tau}{2} A_\alpha\right)^{-1} = E. \tag{3.14}$$

En effet, on a la propriété cherchée en prémultipliant (3.14) par $\left(E + \frac{\tau}{2} A_\alpha\right)^{-1} \left(E - \frac{\tau}{2} A_\alpha\right)$ et en utilisant la commutativité des opérateurs $\left(E - \frac{\tau}{2} A_\alpha\right)$ et $\left(E + \frac{\tau}{2} A_\alpha\right)$ (le dernier résultat est vérifié par multiplication directe).

Ainsi, le problème de stabilité s'est ramené à l'évaluation des normes des opérateurs T_α.

Le lemme de Kellogg appliqué à T_1 et T_2 de (3.13) donne

$$\|T\| \leqslant 1, \tag{3.15}$$

donc

$$\|\psi^{j+1}\| \leqslant \|\psi^j\|. \tag{3.16}$$

On ne doit cependant pas oublier que nous visons à établir la propriété de stabilité du problème initial (3.4). On utilise donc (3.9) et on récrit (3.16)

$$\left\| \left(E + \frac{\tau}{2} A_2\right) \varphi^{j+1} \right\| \leqslant \left\| \left(E + \frac{\tau}{2} A_2\right) \varphi^j \right\|. \tag{3.17}$$

Introduisons la notation

$$\left\| \left(E + \frac{\tau}{2} A_2\right) \varphi \right\| = (C_2 \varphi, \varphi)^{1/2} = \|\varphi\|_{C_2}, \tag{3.18}$$

avec

$$C_\alpha = \left(E + \frac{\tau}{2} A_\alpha^*\right)\left(E + \frac{\tau}{2} A_\alpha\right), \qquad \alpha = 1,\ 2.$$

On voit facilement que $C_\alpha > 0$ et $\|\cdot\|_{C^2}$ est en effet une norme.

Ainsi, on a pour cette norme la condition de stabilité absolue

$$\|\varphi^{j+1}\|_{C_2} \leqslant \|\varphi^j\|_{C_2}. \tag{3.19}$$

On peut donc dire qu'étant donnée une régularité suffisante de la solution de (3.1), si $A_1 \geqslant 0$ et $A_2 \geqslant 0$ et si leurs éléments sont indépendants du temps, le schéma aux différences (3.4) possède la propriété de stabilité absolue et approche le problème initial à l'ordre 2 en τ.

On note enfin que le schéma aux différences de la méthode de stabilisation est facilement réalisable sur ordinateur. En effet, l'équation (3.4) prend la forme

$$\begin{gathered} F^j = A\varphi^j, \qquad \left(E + \frac{\tau}{2} A_1\right)\xi^{j+1/2} = -F^j, \\ \left(E + \frac{\tau}{2} A_2\right)\xi^{j+1} = \xi^{j+1/2}, \quad \varphi^{j+1} = \varphi^j + \tau\xi^{j+1}, \end{gathered} \tag{3.20}$$

$\xi^{j+1/2}$ et ξ^{j+1} étant des quantités auxiliaires qui permettent de réduire (3.4) à plusieurs problèmes simples (3.20). La première et la dernière équation (3.20) sont des relations explicites, i.e. on n'inverse que les opérateurs élémentaires A_1 et A_2.

Considérons le problème non homogène

$$\frac{\partial\varphi}{\partial t} + A\varphi = f, \qquad \varphi = g \quad \text{pour} \quad t = 0, \tag{3.21}$$

où $A = A_1 + A_2$, $A_1 \geqslant 0$, $A_2 \geqslant 0$. Le schéma de stabilisation correspondant s'écrit

$$\left(E + \frac{\tau}{2} A_1\right)\left(E + \frac{\tau}{2} A_2\right)\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^j = f^j, \qquad \varphi^0 = g, \tag{3.22}$$

avec

$$f^j = f(t_{j+1/2}). \tag{3.23}$$

Le problème aux différences (3.22) avec la condition (3.23) approche le problème proposé (3.21) à l'ordre 2 en τ.

Analysons la stabilité du schéma. On transforme (3.22) en

$$\psi^{j+1} = T\psi^j + \tau\left(E + \frac{\tau}{2} A_1\right)^{-1} f^j, \tag{3.24}$$

où

$$\psi^j = \left(E + \frac{\tau}{2} A_2\right)\varphi^j. \tag{3.25}$$

L'équation (3.24) entraîne

$$\| \psi^{j+1} \| \leqslant \| T \| \, \| \psi^j \| + \tau \left\| \left(E + \frac{\tau}{2} A_1 \right)^{-1} \right\| \| f^j \|. \tag{3.26}$$

On a établi

$$\| T \| \leqslant 1$$

pour l'équation homogène, si bien que

$$\| \psi^{j+1} \| \leqslant \| \psi^j \| + \tau \left\| \left(E + \frac{\tau}{2} A_1 \right)^{-1} \right\| \| f^j \|. \tag{3.27}$$

On vérifie de suite que

$$\| f^j \| = \left\| \left(E + \frac{\tau}{2} A_2 \right)^{-1} \left(E + \frac{\tau}{2} A_2 \right) f^j \right\| \leqslant$$

$$\leqslant \left\| \left(E + \frac{\tau}{2} A_2 \right)^{-1} \right\| \left\| \left(E + \frac{\tau}{2} A_2 \right) f^j \right\|. \tag{3.28}$$

Compte tenu de (3.18), (3.25), (3.28), on obtient à partir de (3.27)

$$\| \varphi^{j+1} \|_{C_2} \leqslant \| \varphi^j \|_{C_2} + \tau \left\| \left(E + \frac{\tau}{2} A_1 \right)^{-1} \right\| \left\| \left(E + \frac{\tau}{2} A_2 \right)^{-1} \right\| \| f^j \|_{C_2}. \tag{3.29}$$

On utilise l'estimation

$$\left\| \left(E + \frac{\tau}{2} A_\alpha \right)^{-1} \right\| \leqslant 1 \tag{3.30}$$

établie pour $A_\alpha \geqslant 0$ au Chapitre premier (voir (1.25)) et on a finalement

$$\| \varphi^{j+1} \|_{C_2} \leqslant \| \varphi^j \|_{C_2} + \tau \| f^j \|_{C_2}, \tag{3.31}$$

d'où, de façon récurrente,

$$\| \varphi^j \|_{C_2} \leqslant \| g \|_{C_2} + j\tau \| f \|_{C_2}, \tag{3.32}$$

avec

$$\| f \|_{C_2} = \max_j \| f^j \|_{C_2}. \tag{3.33}$$

On peut donc dire qu'étant donnés $A_1 \geqslant 0$, $A_2 \geqslant 0$ d'éléments indépendants du temps, si la solution φ du problème (3.1) et la fonction f sont suffisamment régulières, il y a stabilité absolue du schéma aux différences (3.22) et l'approximation du problème proposé par (3.22) est d'ordre 2 en τ.

Nous tenons à rappeler une fois de plus que la démonstration donnée n'est valide que pour A indépendant du temps.

5.3.2. Méthode de prédiction-correction

La méthode de prédiction-correction consiste à partager l'intervalle $0 \leqslant t \leqslant T$ en des intervalles partiels et à résoudre le problème (3.1) dans chaque $t_j \leqslant t \leqslant t_{j+1}$ en deux étapes: 1) un schéma

exact à l'ordre 1 possédant une grande « réserve » de stabilité approche le problème au temps $t_{j+1/2} = t_j + \tau/2$ (cette étape est dite *prédiction*), 2) l'équation initiale est approchée à l'ordre 2 sur tout l'intervalle (t_j, t_{j+1}) (*correction*) et on utilise, chose importante, la solution « rude » associée à $t_{j+1/2}$.

Le schéma de prédiction-correction s'écrit donc

$$\begin{aligned} \frac{\varphi^{j+1/4}-\varphi^j}{\tau/2} + A_1\varphi^{j+1/4} &= 0, \\ \frac{\varphi^{j+1/2}-\varphi^{j+1/4}}{\tau/2} + A_2\varphi^{j+1/2} &= 0, \\ \frac{\varphi^{j+1}-\varphi^j}{\tau} + A\varphi^{j+1/2} &= 0, \end{aligned} \tag{3.34}$$

avec la condition $\varphi^0 = g$.

Eliminons la fonction auxiliaire $\varphi^{j+1/4}$ entre les deux premières équations et ramenons le système à

$$\begin{aligned} \left(E+\tfrac{\tau}{2}A_1\right)\left(E+\tfrac{\tau}{2}A_2\right)\varphi^{j+1/2} &= \varphi^j, \\ \frac{\varphi^{j+1}-\varphi^j}{\tau} + A\varphi^{j+1/2} &= 0. \end{aligned} \tag{3.35}$$

On obtient par élimination de $\varphi^{j+1/2}$:

$$\frac{\varphi^{j+1}-\varphi^j}{\tau} + A\left(E+\tfrac{\tau}{2}A_2\right)^{-1}\left(E+\tfrac{\tau}{2}A_1\right)^{-1}\varphi^j = 0, \qquad \varphi^0 = g. \tag{3.36}$$

Voyons le côté approximation. On écrit (3.36) sous la forme

$$\left(E+\tfrac{\tau}{2}A_1\right)\left(E+\tfrac{\tau}{2}A_2\right)\frac{\varphi^{j+1}-\varphi^j}{\tau} + \Lambda\varphi^j = 0,$$

avec

$$\Lambda = \left(E+\tfrac{\tau}{2}A_1\right)\left(E+\tfrac{\tau}{2}A_2\right)A\left(E+\tfrac{\tau}{2}A_2\right)^{-1}\left(E+\tfrac{\tau}{2}A_1\right)^{-1}.$$

On développe le second membre suivant les puissances de τ sous l'hypothèse « $\frac{\tau}{2}\|A_\alpha\| < 1$ » et on obtient

$$\Lambda = A + O(\tau^2).$$

L'évaluation de la méthode de stabilisation permet de dire que la technique de prédiction-correction réalise une approximation d'ordre 2 en τ.

Passons à la stabilité. On récrit (3.36)

$$\left(E+\tfrac{\tau}{2}A_1\right)\left(E+\tfrac{\tau}{2}A_2\right)\frac{\Phi^{j+1}-\Phi^j}{\tau} + A\Phi^j = 0, \tag{3.37}$$

où

$$\Phi^j = \left(E+\tfrac{\tau}{2}A_2\right)^{-1}\left(E+\tfrac{\tau}{2}A_1\right)^{-1}\varphi^j. \tag{3.38}$$

L'équation aux différences (3.37) est stable parce que

$$\| \Phi^{j+1} \|_{C_2} \leqslant \| \Phi^j \|_{C_2}. \tag{3.39}$$

On en tire, compte tenu de (3.38) et (3.18):

$$\left\| \left(E + \frac{\tau}{2} A_1 \right)^{-1} \varphi^{j+1} \right\| \leqslant \left\| \left(E + \frac{\tau}{2} A_1 \right)^{-1} \varphi^j \right\|, \tag{3.40}$$

ou

$$\| \varphi^{j+1} \|_{C_1^{-1}} \leqslant \| \varphi^j \|_{C_1^{-1}}, \tag{3.41}$$

avec

$$C_1^{-1} = \left(E + \frac{\tau}{2} A_1^* \right)^{-1} \left(E + \frac{\tau}{2} A_1 \right)^{-1}.$$

Il y a donc stabilité pour la norme (3.41), ce qui autorise à dire que si $A_1 \geqslant 0$, $A_2 \geqslant 0$, leurs éléments sont indépendants du temps et la solution φ du problème (3.1) présente une régularité suffisante, alors le schéma aux différences (3.34) est absolument stable et approche le problème proposé à l'ordre 2 en τ.

S'agissant d'un problème non homogène, les formules de prédiction-correction sont de la forme

$$\begin{aligned} \frac{\varphi^{j+1/4} - \varphi^j}{\tau/2} + A_1 \varphi^{j+1/4} &= f^j, \\ \frac{\varphi^{j+1/2} - \varphi^{j+1/4}}{\tau/2} + A_2 \varphi^{j+1/2} &= 0, \\ \frac{\varphi^{j+1} - \varphi^j}{\tau} + A \varphi^{j+1/2} &= f^j, \end{aligned} \tag{3.42}$$

où

$$f^j = f(t_{j+1/2}).$$

Avec f^j ainsi choisi, on montre que (3.42) approche le problème initial à l'ordre 2 en τ. Le schéma (3.42) est stable. En effet, on élimine $\varphi^{j+1/2}$ et $\varphi^{j+1/4}$,

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A \left(E + \frac{\tau}{2} A_2 \right)^{-1} \left(E + \frac{\tau}{2} A_1 \right)^{-1} \left(\varphi^j + \frac{\tau}{2} f^j \right) = f^j, \tag{3.43}$$

et on introduit la notation

$$\psi^j = \left(E + \frac{\tau}{2} A_1 \right)^{-1} \left(\varphi^j + \frac{\tau}{2} f^j \right).$$

La relation (3.43) devient

$$\frac{\psi^{j+1} - \psi^j}{\tau} + \left(E + \frac{\tau}{2} A_1 \right)^{-1} A \left(E + \frac{\tau}{2} A_2 \right)^{-1} \psi^j = \left(E + \frac{\tau}{2} A_1 \right)^{-1} \left(\frac{f^j + f^{j+1}}{2} \right).$$

D'où

$$\psi^{j+1}=\left[E-\left(E+\frac{\tau}{2}A_1\right)^{-1}\tau A\left(E+\frac{\tau}{2}A_2\right)^{-1}\right]\psi^j+$$
$$+\tau\left(E+\frac{\tau}{2}A_1\right)^{-1}\left(\frac{f^j+f^{j+1}}{2}\right). \qquad (3.44)$$

Puisque

$$E-\left(E+\frac{\tau}{2}A_1\right)^{-1}\tau A\left(E+\frac{\tau}{2}A_2\right)^{-1}=$$
$$=\left(E+\frac{\tau}{2}A_1\right)^{-1}\left[\left(E+\frac{\tau}{2}A_1\right)\left(E+\frac{\tau}{2}A_2\right)-\tau A\right]\left(E+\frac{\tau}{2}A_2\right)^{-1}=$$
$$=\left(E+\frac{\tau}{2}A_1\right)^{-1}\left[\left(E-\frac{\tau}{2}A_1\right)\left(E-\frac{\tau}{2}A_2\right)\right]\left(E+\frac{\tau}{2}A_2\right)^{-1}=$$
$$=\left[\left(E+\frac{\tau}{2}A_1\right)^{-1}\left(E-\frac{\tau}{2}A_1\right)\right]\left[\left(E-\frac{\tau}{2}A_2\right)\left(E+\frac{\tau}{2}A_2\right)^{-1}\right],$$

le lemme de Kellogg et (1.25) du Chapitre premier entraînent

$$\left\|\varphi^j+\frac{\tau}{2}f^j\right\|_{C_1^{-1}}\leqslant\left\|g+\frac{\tau}{2}f^0\right\|_{C_1^{-1}}+\tau j\left\|f\right\|_{C_1^{-1}}, \qquad (3.45)$$

avec $\|f\|_{C_1^{-1}}=\max\limits_j\|f^j\|_{C_1^{-1}}$, i.e. pour $0\leqslant t_j\leqslant T$, on a le résultat cherché.

Ainsi, le schéma aux différences (3.42) est absolument stable et donne une solution exacte à l'ordre 2 en τ à condition que $A_1\geqslant 0$, $A_2\geqslant 0$ soient des matrices d'éléments indépendants du temps et la solution et le second membre f du problème (3.1) présentent une régularité suffisante.

Une dernière remarque. On montre que la stabilité absolue établie de (3.34) n'exclut pas l'instabilité absolue de son schéma correcteur pris séparément. Pour simplifier l'exposé, on se place dans le cas où A est le laplacien discrétisé en dimension 2 et la solution s'annule sur la frontière du carré unité D.

Le correcteur s'écrit

$$\frac{\varphi_{k,l}^{j+1}-\varphi_{k,l}^{j-1}}{2\tau}-\frac{\varphi_{k+1,l}^{j}-2\varphi_{k,l}^{j}+\varphi_{k-1,l}^{j}}{h^2}-\frac{\varphi_{k,l+1}^{j}-2\varphi_{k,l}^{j}+\varphi_{k,l-1}^{j}}{h^2}=0$$

(comme on n'a pas recours au prédicteur, seuls j entiers sont considérés).

On trouve une solution φ^j de ce problème aux différences, qui ne vérifie pas la condition (3.16).

On cherche la solution sous la forme

$$\varphi_{k,l}^{j}=\lambda^j\sin m\pi kh\cdot\sin p\pi lh,$$

j du premier membre étant un indice et j du second l'exposant.

Portons cette expression dans l'équation aux différences, il vient l'équation caractéristique

$$\lambda^2 + 8a_{mp}\lambda - 1 = 0,$$

où

$$a_{mp} = \frac{\tau}{h^2}\left(\sin^2\frac{m\pi h}{2} + \sin^2\frac{p\pi h}{2}\right).$$

Si l'on prend

$$\lambda = -4a_{mp} - \sqrt{1 + 16a_{mp}^2},$$

on obtient quand $\tau \to 0$ de façon que $\tau/h^2 = \text{const}$:

$$\| \varphi^j \| / \| \varphi^0 \| = |\lambda|^j \to \infty,$$

i.e. il y a instabilité absolue.

On voit que le schéma de correction a beau être absolument instable : la « réserve » de stabilité du prédicteur garantit la stabilité absolue globale.

5.3.3. Méthode de splitting-up

Les procédés de stabilisation et de prédiction-correction sont équivalents du point de vue de la précision et possèdent la stabilité absolue si $A_\alpha \geqslant 0$.

Les opérateurs A_α ont été assujettis à être indépendants de t. Cette contrainte nous a permis d'établir les propriétés de stabilité sous la seule hypothèse constructive « A_α sont semi-définis positifs ». Avec A_α dépendant du temps, cette analyse de la stabilité n'aboutit en général pas. La méthode de splitting-up constitue une exception à la règle.

Soit, dans (3.1)-(3.3), $A_1(t) \geqslant 0$ et $A_2(t) \geqslant 0$. Considérons leurs approximations sur l'intervalle $t_j \leqslant t \leqslant t_{j+1}$

$$\Lambda_\alpha^j = A_\alpha(t_{j+1/2}),$$

les éléments de ces matrices étant supposés suffisamment réguliers. En discrétisant selon les schémas de Crank-Nicholson simples, on obtient le système (proposé par Yanenko [XV; 13]) *)

$$\begin{aligned} \frac{\varphi^{j+1/2} - \varphi^j}{\tau} + \Lambda_1^j \frac{\varphi^{j+1/2} + \varphi^j}{2} &= 0, \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + \Lambda_2^j \frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} &= 0. \end{aligned} \tag{3.46}$$

Le système (3.46) se ramène par élimination des fonctions auxiliaires $\varphi^{j+1/2}$ à l'équation

$$\varphi^{j+1} = T^j \varphi^j, \tag{3.47}$$

*) Pour la théorie et les variantes voir le rapport de l'Auteur au Second Symposium on numerical solution of partial differential equations tenu en 1970 aux Etats-Unis (SYNSPADE 1970).

où

$$T^j = \left(E + \frac{\tau}{2}\Lambda_2^j\right)^{-1}\left(E - \frac{\tau}{2}\Lambda_2^j\right)\left(E + \frac{\tau}{2}\Lambda_1^j\right)^{-1}\left(E - \frac{\tau}{2}\Lambda_1^j\right). \tag{3.48}$$

On commence par le côté approximation. On développe l'opérateur T^j suivant les puissances de τ dans la condition $\frac{\tau}{2} \| \Lambda_\alpha^j \| < 1$. Des transformations simples aboutissent à

$$T^j = E - \tau\Lambda^j + \frac{\tau^2}{2}((\Lambda_1^j)^2 + 2\Lambda_2^j\Lambda_1^j + (\Lambda_2^j)^2) - \ \ldots \tag{3.49}$$

Si Λ_α^j commutent, i.e. $\Lambda_1^j\Lambda_2^j = \Lambda_2^j\Lambda_1^j$, alors (3.49) se récrit

$$T = E - \tau\Lambda^j + \frac{\tau^2}{2}(\Lambda^j)^2 - \ \ldots \tag{3.50}$$

Si $A_1(t) \geqslant 0$, $A_2(t) \geqslant 0$ et si les éléments de ces matrices et la solution φ du problème (3.1)-(3.3) accusent une régularité suffisante, le schéma aux différences (3.46) est donc absolument stable (cette propriété découle de suite de l'inégalité $\| T^j \| < 1$ qui est valable par Kellogg) et il approche (3.1) à l'ordre 2 en τ pour Λ_1^j et Λ_2^j commutatifs et à l'ordre 1 dans le cas contraire.

On approche maintenant $A_1(t)$ et $A_2(t)$ non pas sur $t_j \leqslant t \leqslant t_{j+1}$ (cas (3.46)), mais sur $t_{j-1} \leqslant t \leqslant t_{j+1}$. On pose

$$\Lambda_\alpha^j = A_\alpha(t_j).$$

Soit deux systèmes aux différences

$$\begin{aligned} \frac{\varphi^{j-1/2} - \varphi^{j-1}}{\tau} + \Lambda_1^j \frac{\varphi^{j-1/2} + \varphi^{j-1}}{2} = 0, \\ \frac{\varphi^j - \varphi^{j-1/2}}{\tau} + \Lambda_2^j \frac{\varphi^j + \varphi^{j-1/2}}{2} = 0 \end{aligned} \tag{3.51}$$

et

$$\begin{aligned} \frac{\varphi^{j+1/2} - \varphi^j}{\tau} + \Lambda_2^j \frac{\varphi^{j+1/2} + \varphi^j}{2} = 0, \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + \Lambda_1^j \frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} = 0. \end{aligned} \tag{3.52}$$

Le cycle de calcul consiste justement à appliquer à tour de rôle (3.51) et (3.52). On procède comme plus haut et on montre que le cycle complet aboutit dans ce cas à

$$\varphi^{j+1} = T^j\varphi^{j-1}, \tag{3.53}$$

où

$$T^j = \left(E + \frac{\tau}{2}\Lambda_1^j\right)^{-1}\left(E - \frac{\tau}{2}\Lambda_1^j\right)\left(E + \frac{\tau}{2}\Lambda_2^j\right)^{-1}\left(E - \frac{\tau}{2}\Lambda_2^j\right)\times$$

$$\times\left(E + \frac{\tau}{2}\Lambda_2^j\right)^{-1}\left(E - \frac{\tau}{2}\Lambda_2^j\right)\left(E + \frac{\tau}{2}\Lambda_1^j\right)^{-1}\left(E - \frac{\tau}{2}\Lambda_1^j\right) =$$

$$= E - 2\tau\Lambda^j + \frac{(2\tau)^2}{2}(\Lambda^j)^2 - \ldots$$

Comparons T^j et son homologue du schéma de Crank-Nicholson

$$\frac{\varphi^{j+1} - \varphi^{j-1}}{2\tau} + \Lambda^j \frac{\varphi^{j+1} + \varphi^{j-1}}{2} = 0.$$

On constate que T^j pour le schéma de décomposition 2-cyclique et l'opérateur du schéma de Crank-Nicholson associé à l'intervalle temporel double coïncident à $O(\tau^2)$ près que les opérateurs A_α commutent ou non. Cet artifice permet donc de négliger cette contrainte sévère qu'est la commutativité des opérateurs.

Pour ce qui est de la stabilité numérique, on considère la relation (3.47) qui entraîne

$$\| \varphi^{j+1} \| \leqslant \| T^j \| \| \varphi^j \|.$$

On a

$$\| T^j \| \leqslant 1$$

pour $A_\alpha \geqslant 0$ (voir plus haut), si bien que

$$\| \varphi^{j+1} \| \leqslant \| \varphi^j \|. \tag{3.54}$$

Il en résulte de suite

$$\| \varphi^j \| \leqslant \| g \|. \tag{3.55}$$

S'agissant de la méthode 2-cyclique, on a à chaque pas une estimation de la forme (3.54), ce qui signifie la stabilité absolue. (Un procédé analogue de symétrisation a été utilisé par Strang [VII], Samarski [III; 12] et Friazinov [XV; 12].)

On dit par conséquent qu'étant données $A_1(t) \geqslant 0$, $A_2(t) \geqslant 0$, si la solution φ du problème (3.1)-(3.3) et les éléments des matrices $A_1(t)$, $A_2(t)$ sont suffisamment réguliers, le système aux différences (3.51), (3.52) possède la propriété de stabilité absolue et le schéma (3.53) approche (3.1) à l'ordre 2 en τ.

On aborde le problème non homogène par la décomposition complète 2-cyclique. On considère à cet effet un système (3.51), (3.52)

mis sous une forme plus commode, à savoir :

$$\begin{aligned}
\left(E+\frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-1/2}&=\left(E-\frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-1},\\
\left(E+\frac{\tau}{2}\Lambda_2^j\right)(\varphi^j-\tau f^j)&=\left(E-\frac{\tau}{2}\Lambda_2^j\right)\varphi^{j-1/2},\\
\left(E+\frac{\tau}{2}\Lambda_2^j\right)\varphi^{j+1/2}&=\left(E-\frac{\tau}{2}\Lambda_2^j\right)(\varphi^j+\tau f^j),\\
\left(E+\frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+1}&=\left(E-\frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+1/2},
\end{aligned} \tag{3.56}$$

où $f^j = f(t_j)$. Résolvons par rapport à φ^{j+1} :

$$\varphi^{j+1}=T^j\varphi^{j-1}+2\tau T_1^jT_2^jf^j, \tag{3.57}$$

avec

$$T^j=T_1^jT_2^jT_2^jT_1^j, \tag{3.58}$$

$$T_\alpha^j=\left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)^{-1}\left(E-\frac{\tau}{2}\Lambda_\alpha^j\right). \tag{3.59}$$

On développe (3.57) suivant les puissances du petit paramètre τ et on obtient

$$\varphi^{j+1}=\left[E-2\tau\Lambda^j+\frac{(2\tau)^2}{2}(\Lambda^j)^2\right]\varphi^{j-1}+2\tau(E-\tau\Lambda^j)f^j+O(\tau)^3, \tag{3.60}$$

qui devient par certaines transformations

$$\frac{\varphi^{j+1}-\varphi^{j-1}}{2\tau}+\Lambda^j(E-\tau\Lambda^j)\varphi^{j-1}=(E-\tau\Lambda^j)f^j+O(\tau^2). \tag{3.61}$$

On élimine φ^{j-1} de la dernière relation en développant la solution en série de Taylor autour du point t_{j-1}. On a à τ^2 près

$$\varphi^j=\varphi^{j-1}+\left(\frac{\partial\varphi}{\partial t}\right)^{j-1}\tau+O(\tau^2). \tag{3.62}$$

On élimine la dérivée $\frac{\partial\varphi}{\partial t}$ à l'aide de la relation

$$\left(\frac{\partial\varphi}{\partial t}\right)^{j-1}=-\Lambda^j\varphi^{j-1}+f^j+O(\tau). \tag{3.63}$$

Portons (3.63) dans (3.62), il vient

$$\varphi^j=(E-\tau\Lambda^j)\varphi^{j-1}+\tau f^j+O(\tau^2),$$

d'où

$$(E-\tau\Lambda^j)\varphi^{j-1}=\varphi^j-\tau f^j+O(\tau^2). \tag{3.64}$$

On porte dans (3.61), ce qui donne

$$\frac{\varphi^{j+1}-\varphi^{j-1}}{2\tau}+\Lambda^j\varphi^j=f^j+O(\tau^2). \tag{3.65}$$

L'équation (3.65) approche évidemment (3.1) sur $t_{j-1} \leqslant t \leqslant t_{j+1}$ à l'ordre 2 en τ. Ainsi, la méthode 2-cyclique conduit à une approximation aux différences d'ordre 2 de l'équation d'évolution.

La démonstration de la stabilité pour la norme énergie s'effectue de manière élémentaire. En effet, on évalue (3.57) en norme:

$$\| \varphi^{j+1} \| \leqslant \| T^j \| \| \varphi^{j-1} \| + 2\tau \| T_1^j \| \| T_2^j \| \| f^j \|. \tag{3.66}$$

Il est connu que $\| T_\alpha^j \| \leqslant 1$, donc

$$\| T^j \| \leqslant \| T_1^j \| \| T_2^j \| \| T_2^j \| \| T_1^j \| \leqslant 1.$$

Aussi

$$\| \varphi^{j+1} \| \leqslant \| \varphi^{j+1} \| + 2\tau \| f^j \|. \tag{3.67}$$

On obtient moyennant la relation de récurrence (3.63)

$$\| \varphi^j \| \leqslant \| g \| + \tau j \| f \|, \tag{3.68}$$

avec

$$\| f \| = \max_j \| f^j \|.$$

L'estimation (3.68) implique la stabilité numérique du schéma sur tout intervalle temporel fini.

Le système (3.56) se met sous forme équivalente

$$\begin{aligned}
\left(E + \tfrac{\tau}{2} \Lambda_1^j\right) \varphi^{j-2/3} &= \left(E - \tfrac{\tau}{2} \Lambda_1^j\right) \varphi^{j-1}, \\
\left(E + \tfrac{\tau}{2} \Lambda_2^j\right) \varphi^{j-1/3} &= \left(E - \tfrac{\tau}{2} \Lambda_2^j\right) \varphi^{j-2/3}, \\
\varphi^{j+1/3} &= \varphi^{j-1/3} + 2\tau f^j, \\
\left(E + \tfrac{\tau}{2} \Lambda_2^j\right) \varphi^{j+2/3} &= \left(E - \tfrac{\tau}{2} \Lambda_2^j\right) \varphi^{j+1/3}, \\
\left(E + \tfrac{\tau}{2} \Lambda_1^j\right) \varphi^{j+1} &= \left(E - \tfrac{\tau}{2} \Lambda_1^j\right) \varphi^{j+2/3}.
\end{aligned} \tag{3.69}$$

L'élimination des inconnues munies d'indices fractionnaires donne l'équation résolue

$$\varphi^{j+1} = T_1^j T_2^j T_2^j T_1^j \varphi^{j-1} + 2\tau T_1^j T_2^j f^j \tag{3.70}$$

qui coïncide avec (3.57). L'écriture (3.69) est des fois à préférer à (3.56).

On énonce donc le résultat suivant: si $A_1(t) \geqslant 0$, $A_2(t) \geqslant 0$ et si la solution φ, la fonction $f(t)$ et les éléments des matrices mentionnées sont suffisamment réguliers, alors il y a stabilité absolue sur $0 \leqslant t \leqslant T$ du système d'équations aux différences (3.56) et ce schéma approche l'équation proposée à l'ordre 2 en τ.

5.3.4. Certaines remarques générales

On compare les méthodes de décomposition de ce numéro sous les hypothèses « A, A_1, A_2 sont indépendants du temps; $A_1 \geqslant 0$, $A_2 \geqslant 0$ et $A_1A_2 = A_2A_1$ ». Ce cas simple a attiré l'attention de nombreux auteurs. Ecrivons les schémas aux différences qui sont formellement résolus à chaque pas par rapport à la solution cherchée. On constate sans peine l'équivalence de tous les schémas de décomposition pour le problème évolutif (3.1) avec $f = 0$ qui ne sont donc qu'autant de schémas de mise en œuvre. On les ramène en particulier à la forme

$$\varphi^{j+1} = T\varphi^j,$$

où

$$T = \left(E + \frac{\tau}{2}A_1\right)^{-1}\left(E - \frac{\tau}{2}A_1\right)\left(E + \frac{\tau}{2}A_2\right)^{-1}\left(E - \frac{\tau}{2}A_2\right).$$

S'agissant du problème non homogène, les schémas de décomposition ne sont équivalents que pour ce qui est de l'ordre d'approximation, i.e. divers schémas approchant à l'ordre 2 en τ conduisent avec une erreur en τ^2 aux résultats différant d'une quantité en $O(\tau^2)$.

La deuxième remarque a trait aux schémas de décomposition pour $A_1 \geqslant 0$, $A_2 \geqslant 0$ indépendants du temps tels que $A_1A_2 \neq A_2A_1$. On approche la solution des problèmes d'évolution à l'aide de tous les schémas considérés : stabilisation, prédiction-correction, splitting-up. Ces méthodes sont équivalentes pour ce qui est de l'ordre de précision, mais on ne saurait les identifier parce que l'opérateur de transition T diffère d'un procédé à l'autre. La question des sphères d'influence des schémas reste ouverte, encore que le fait de pouvoir aborder un même cas par trois méthodes différentes (autonomes) encourage un chercheur qui doit faire face à des problèmes complexes.

Nous voulons enfin parler du cas plus général avec $A_1 \geqslant 0$, $A_2 \geqslant 0$, $A_1A_2 \neq A_2A_1$, les opérateurs A_1 et A_2 dépendant de t. On s'adresse de préférence à la technique de splitting-up dont la variante 2-cyclique donne une solution exacte à l'ordre 2 en τ.

5.4. Méthodes de décomposition pour $n > 2$

Jusqu'à présent l'opérateur A a été représenté par la somme de deux opérateurs plus simples. Le cas avec un nombre quelconque d'opérateurs intervient souvent en physique mathématique. Soit

$$A = \sum_{\alpha=1}^{n} A_\alpha, \tag{4.1}$$

$A_\alpha \geqslant 0$. Le cas $n = 2$ étant étudié au n° 5.3, on s'intéresse seulement à $n > 2$.

On constate en premier lieu que les méthodes de décomposition avec $n = 2$ ne se transposent pas trivialement au cas $n > 2$.

5.4.1. Méthode de stabilisation

La méthode de stabilisation s'écrit sous l'hypothèse (4.1)

$$\prod_{\alpha=1}^{n}\left(E+\frac{\tau}{2}A_\alpha\right)\frac{\varphi^{j+1}-\varphi^j}{\tau}+A\varphi^j=f^j,\quad \varphi^0=g, \tag{4.2}$$

où

$$f^j=f(t_{j+1/2}).$$

L'algorithme procède comme suit:

$$\begin{aligned}
F^j &= -A\varphi^j+f^j,\\
\left(E+\frac{\tau}{2}A_1\right)\xi^{j+1/n} &= F^j,\\
\left(E+\frac{\tau}{2}A_2\right)\xi^{j+2/n} &= \xi^{j+1/n},\\
&\cdots\cdots\cdots\cdots\\
\left(E+\frac{\tau}{2}A_n\right)\xi^{j+1} &= \xi^{j+\frac{n-1}{n}},\\
\varphi^{j+1} &= \varphi^j+\tau\xi^{j+1}.
\end{aligned} \tag{4.3}$$

Il est immédiat de vérifier que la méthode est exacte à l'ordre 2 en τ s'il y a régularité suffisante de la solution. La stabilité numérique est définie par la condition

$$\| T \| \leqslant 1, \tag{4.4}$$

T étant l'opérateur de transition

$$T=E-\tau\prod_{\alpha=n}^{1}\left(E+\frac{\tau}{2}A_\alpha\right)^{-1}A. \tag{4.5}$$

Contrairement au cas $n = 2$, la condition $A_\alpha \geqslant 0$ n'entraîne pas la stabilité pour une norme.

On établit le résultat de stabilité par un artifice algorithmique simple, à savoir on pose $f^j = 0$ et on ramène (4.2) résolue par rapport à φ^{j+1} à la forme

$$\varphi^{j+1} = T\varphi^j. \tag{4.6}$$

L'opérateur T étant supposé indépendant du temps (i.e. de l'indice j), on s'intéresse, en résolvant le problème (4.6) avec la condition initiale

$$\varphi^0 = g \tag{4.7}$$

et τ fixe garantissant l'approximation nécessaire, à l'allure de la norme $\| \varphi^j \|$. Si $\| \varphi^j \|$ ne croît pas, on a visiblement $\| T \| < 1$,

i.e. on est dans la condition de stabilité numérique. On passe au problème non homogène. L'équation (4.2) se récrit

$$\varphi^{j+1} = T\varphi^j + \tau \prod_{\alpha=n}^{1} \left(E + \frac{\tau}{2} A_\alpha\right)^{-1} f^j. \tag{4.8}$$

D'où

$$\| \varphi^{j+1} \| \leqslant \| T \| \, \| \varphi^j \| + \tau \sum_{\alpha=n}^{1} \left\| \left(E + \frac{\tau}{2} A_\alpha\right)^{-1} \right\| \| f^j \|,$$

ou

$$\| \varphi^{j+1} \| \leqslant \| \varphi^j \| + \tau \| f^j \| \tag{4.9}$$

par suite de (4.4) et de l'inégalité (1.25) du Chapitre premier. On aboutit par récurrence à la condition de stabilité

$$\| \varphi \| \leqslant \| g \| + \tau j \| f \|, \tag{4.10}$$

où

$$\| f \| = \max_j \| f^j \|. \tag{4.11}$$

Il n'est nullement nécessaire d'utiliser (4.7) pour vérifier la condition de stabilité $\| T \| \leqslant 1$ de (4.6). On peut prendre comme condition initiale n'importe quelle fonction de même classe de régularité que g et f.

5.4.2. Méthode de prédiction-correction

Le schéma de décomposition s'écrit

$$\begin{aligned}
&\left(E + \frac{\tau}{2} A_1\right) \varphi^{j+\frac{1}{2n}} = \varphi^j + \frac{\tau}{2} f^j, \\
&\left(E + \frac{\tau}{2} A_2\right) \varphi^{j+\frac{2}{2n}} = \varphi^{j+\frac{1}{2n}}, \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
&\left(E + \frac{\tau}{2} A_n\right) \varphi^{j+1/2} = \varphi^{j+\frac{n-1}{2n}}, \\
&\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^{j+1/2} = f^j,
\end{aligned} \tag{4.12}$$

où $A_\alpha \geqslant 0$ et $f^j = f(t_{j+1/2})$. Le système (4.12) se réduit à une équation de la forme

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A \prod_{\alpha=n}^{1} \left(E + \frac{\tau}{2} A_\alpha\right)^{-1} \left(\varphi^j + \frac{\tau}{2} f^j\right) = f^j, \tag{4.13}$$

avec la condition

$$\varphi^0 = g.$$

Si la solution est suffisamment régulière, la méthode est exacte à l'ordre 2 en τ. L'équation (4.13) se récrit

$$\varphi^{j+1} = T\varphi^j + \frac{\tau}{2}(E+T)f^j, \tag{4.14}$$

avec T l'opérateur de transition défini par la formule

$$T = E - \tau A \prod_{\alpha=n}^{1} \left(E + \frac{\tau}{2} A_\alpha\right)^{-1}. \tag{4.15}$$

Le critère de stabilité numérique se ramène en fin de compte à évaluer la norme de T. La question reste ouverte car la condition constructive $A_\alpha \geqslant 0$ ne suffit malheureusement pas pour démontrer la stabilité du schéma.

Pour terminer l'analyse des deux schémas considérés, plaçons-nous dans le cas simple où A_α commutent et possèdent une base commune. Cette condition supplémentaire s'avère suffisante pour que $A_\alpha \geqslant 0$ détermine la stabilité. En effet, s'il y a commutativité, les deux schémas ont le même T. Considérons, pour simplifier l'exposé, le problème homogène (4.6), (4.7) dont la solution sera cherchée sous forme spectrale

$$\varphi^j = \sum_k \varphi_k^j u_k, \tag{4.16}$$

u_k étant les fonctions propres du problème (1.7) (voir chap. 1) et $\varphi_k^j = (\varphi^j, u_k^*)$, u_k^* étant les fonctions propres du problème adjoint (1.7) du n° 1.1. Puisque $\{u_k\}$ est une base commune, on a

$$Au_k = \lambda_k u_k, \quad A_\alpha u_k = \lambda_k^\alpha u_k, \quad \lambda_k = \sum_{\alpha=1}^{n} \lambda_k^\alpha. \tag{4.17}$$

Portons (4.16) et l'expression correspondante de g dans (4.6), (4.7), il vient pour les coefficients de Fourier φ_k^j

$$\varphi_k^{j+1} = T_k \varphi_k^j, \qquad \varphi_k^0 = g_k, \tag{4.18}$$

où

$$T_k = 1 - \frac{\tau\lambda_k}{\prod_{\alpha=1}^{n} \left(1 + \frac{\tau}{2}\lambda_k^\alpha\right)}. \tag{4.19}$$

On transforme (4.19) en

$$T_k = \frac{\mu_k - \frac{\tau}{2}\lambda_k}{\mu_k + \frac{\tau}{2}\lambda_k}, \tag{4.20}$$

μ_k étant des constantes positives si $\lambda_k^\alpha \geqslant 0$. Il résulte de (4.20)

$$| T_k | \leqslant 1, \tag{4.21}$$

d'où la stabilité (voir n° 1.3).

La méthode de stabilisation et la prédiction-correction s'appliquent également à A dépendant du temps encore que établir à priori la condition de stabilité s'avère dans ce cas un problème plus délicat. On hésite donc à dire s'il y a lieu d'appliquer les deux schémas dans des situations générales. Cette circonstance a incité l'auteur à formuler un procédé de décomposition plus ou moins « passe-partout » pour des problèmes complexes et assez généraux. Il s'agit de la méthode de splitting-up successif 2-cyclique.

5.4.3. Méthode de splitting-up en termes de schémas élémentaires

On se propose de construire un problème discrétisé approché d'ordre 2 en τ qui soit absolument stable dans le temps. Conformément à l'hypothèse de $n > 2$, on suppose que

$$\Lambda^j = \sum_{\alpha=1}^{n} \Lambda_\alpha^j, \tag{4.22}$$

tous les Λ_α^j étant des opérateurs semi-définis positifs: $\Lambda_\alpha^j \geqslant 0$. Soit le système d'équations

$$\left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+\frac{\alpha}{n}} = \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+\frac{\alpha-1}{n}}, \quad \alpha = 1,2,\ldots, n. \tag{4.23}$$

Si $\Lambda_\alpha^j \geqslant 0$ commutent et $\Lambda_\alpha^j = A_\alpha^{j+1/2}$ ou $\Lambda_\alpha^j = \frac{1}{2}(A^{j+1} + A^j)$, le schéma (4.23) est inconditionnellement stable et constitue une approximation d'ordre 2. On l'établit sans trop de difficulté à l'aide de la méthode de Fourier. S'agissant de Λ_α^j non commutatifs, le schéma est en général, on le voit aisément, exact à l'ordre 1 en τ. On comprend donc qu'il est moins intéressant pour les applications que le schéma d'ordre d'approximation 2 proposé par Diakonov [XV; 5]:

$$\begin{gathered}\Phi^{j+\frac{\alpha}{2n}} = \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+\frac{\alpha-1}{2n}}, \qquad \alpha = 1,\ 2,\ \ldots,\ n, \\ \left(E + \frac{\tau}{2}\Lambda_{2n-\alpha+1}^j\right)\Phi^{j+\frac{\alpha}{2n}} = \Phi^{j+\frac{\alpha-1}{2n}}, \\ \alpha = n+1,\ n+2,\ \ldots,\ 2n.\end{gathered} \tag{4.24}$$

Etablissons les caractéristiques de la décomposition complète par le schéma (4.23) qui fournit la solution du problème de Cauchy

pour Λ_α^j non commutatifs et semi-définis positifs et est d'ordre d'approximation 2. Cela équivaut, dans un sens, à résoudre complètement le problème de décomposition.

Le système (4.23) se ramène à une équation

$$\Phi^{j+1} = \prod_{\alpha=1}^{n} \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right) \Phi^j. \tag{4.25}$$

L'égalité (4.25) aidant, on établit la majoration en norme

$$\| \Phi^{j+1} \| \leqslant \prod_{\alpha=1}^{n} \left\| \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right) \right\| \| \Phi^j \|. \tag{4.26}$$

D'après Kellogg,

$$\| \Phi^{j+1} \| \leqslant \| \Phi^j \| \leqslant \ldots \leqslant \| g \|. \tag{4.27}$$

Dans le cas d'opérateurs symétriques gauches $((\Lambda_\alpha^j \varphi, \varphi) = 0)$,

$$\| \Phi^{j+1} \| = \| \Phi^j \| = \ldots = \| g \|. \tag{4.28}$$

Ainsi, le schéma est absolument stable.

Quel est son ordre d'approximation? Développons

$$T^j = \prod_{\alpha=1}^{n} \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right)$$

$\left(\text{en posant } \frac{\tau}{2} \| \Lambda_\alpha \| < 1\right)$ suivant les puissances de τ petit. Comme

$$T^j = \prod_{\alpha=1}^{n} T_\alpha^j,$$

on commence par développer T_α^j:

$$T_\alpha^j = E - \tau\Lambda_\alpha^j + \frac{\tau^2}{2}(\Lambda_\alpha^j)^2 \ldots \tag{4.29}$$

Finalement

$$T^j = E - \tau\Lambda^j + \frac{\tau^2}{2}\left[(\Lambda^j)^2 + \sum_{\alpha=1}^{n} \sum_{\beta=\alpha+1}^{n} (\Lambda_\alpha^j \Lambda_\beta^j - \Lambda_\beta^j \Lambda_\alpha^j)\right] + O(\tau^3). \tag{4.30}$$

Avec Λ_α^j commutatifs, l'expression sous le signe $\sum\sum$ s'annule et on a

$$T^j = E - \tau\Lambda^j + \frac{\tau^2}{2}(\Lambda^j)^2 + O(\tau^3). \tag{4.31}$$

On compare (4.31) avec le développement de T^j de (1.8) en série suivant les puissances de $\tau\Lambda^j$, Λ^j étant défini moyennant (1.22)-(1.24), et on voit que dans ce cas particulier (4.23) approche à l'ordre 2 en τ. Si Λ_α^j ne commutent pas, il est exact à l'ordre 1 en τ. On obtient un schéma précis à l'ordre 2 en remplaçant (4.23) par

$$\Phi^j = \prod_{\alpha=1}^{n} T_\alpha^j \Phi^{j-1}, \qquad \Phi^{j+1} = \prod_{\alpha=n}^{1} T_\alpha^j \Phi^j. \tag{4.32}$$

Algorithmiquement, cela signifie qu'on résout d'abord le système (4.23) sur l'intervalle $t_{j-1} \leqslant t \leqslant t_j$ pour $\alpha = 1, 2, \ldots, n$, puis le même système sur $t_j \leqslant t \leqslant t_{j+1}$ pour $\alpha = n, n-1, \ldots, 1$:

$$\begin{gathered}\left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+\frac{\alpha}{n}-1} = \left(E-\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+\frac{\alpha-1}{n}-1} \\ \alpha = 1, 2, \ldots, n, \\ \left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+1-\frac{\alpha-1}{n}} = \left(E-\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+1-\frac{\alpha}{n}}, \\ \alpha = n, n-1, \ldots, 1.\end{gathered} \tag{4.33}$$

On a évidemment pour le cycle complet (4.33)

$$\Phi^{j+1} = T^j \Phi^{j-1},$$

où

$$T^j = \prod_{\alpha=1}^{n} T_\alpha^j \prod_{\alpha=n}^{1} T_\alpha^j = E - 2\tau\Lambda^j + \frac{(2\tau)^2}{2}(\Lambda^j)^2 + O(\tau^3).$$

Ainsi, le schéma (4.33) est exact à l'ordre 2 en τ sur l'intervalle $t_{j-1} \leqslant t \leqslant t_{j+1}$ à condition de prendre pour Λ_α^j l'un des opérateurs de (1.22)-(1.24).

On note encore la stabilité absolue de ce schéma pour $\Lambda_\alpha^j \geqslant 0$.

On associe à l'équation non homogène

$$\begin{gathered}\frac{\partial\varphi}{\partial t} + A\varphi = f, \\ \varphi = g \text{ pour } t = 0,\end{gathered} \tag{4.34}$$

где $A(t) \geqslant 0$ и $A = \sum_{\alpha=1}^{n} A_\alpha$, $A_\alpha(t) \geqslant 0$ на $t_{j-1} \leqslant t \leqslant t_{j+1}$, схема декомпозиции

$$\begin{aligned}
&\left(E + \frac{\tau}{2}\Lambda_1^j\right)\Phi^{j-\frac{n-1}{n}} = \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-1},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\left(E + \frac{\tau}{2}\Lambda_n^j\right)(\varphi^j - \tau f^j) = \left(E - \frac{\tau}{2}\Lambda_n^j\right)\varphi^{j-\frac{1}{n}},\\
&\left(E + \frac{\tau}{2}\Lambda_n^j\right)\varphi^{j+\frac{1}{n}} = \left(E - \frac{\tau}{2}\Lambda_n^j\right)(\varphi^j + \tau f^j),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\left(E + \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+1} = \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+\frac{n-1}{n}},
\end{aligned} \tag{4.35}$$

где

$$\Lambda_\alpha^j = A_\alpha(t_j).$$

Легко устанавливается, что она имеет порядок аппроксимации 2 по τ и абсолютно устойчива при φ, предполагаемом достаточно регулярным.

Как и при $n = 2$, система (4.35) эквивалентна

$$\begin{aligned}
&\left(E + \frac{\tau}{2}\Lambda_\alpha\right)\Phi^{j-\frac{(n+1)-\alpha}{n+1}} = \left(E - \frac{\tau}{2}\Lambda_\alpha\right)\Phi^{j-\frac{(n+1)-\alpha+1}{(n+1)}},\\
&\alpha = 1, 2, \ldots, n,\\
&\Phi^{j+\frac{1}{(n+1)}} = \Phi^{j-\frac{1}{(n+1)}} + 2\tau f^j,\\
&\left(E + \frac{\tau}{2}\Lambda_{n-\alpha+2}\right)\Phi^{j+\frac{\alpha}{(n+1)}} = \left(E - \frac{\tau}{2}\Lambda_{n-\alpha+2}\right)\Phi^{j+\frac{\alpha-1}{(n+1)}},\\
&\alpha = 2, 3, \ldots, n+1.
\end{aligned} \tag{4.36}$$

Перейдем теперь к случаю неявных разностных аппроксимаций и рассмотрим задачу

$$\begin{aligned}
&\frac{\partial\varphi}{\partial t} + A\varphi = 0 \quad \text{в } D \times D_t,\\
&\varphi = g^{*}) \quad \text{в } D \quad \text{при } t = 0.
\end{aligned} \tag{4.37}$$

*) Рассматриваются однородные граничные условия по соображениям методического характера.

On suppose que $A = \sum_{\alpha=1}^{n} A_\alpha$, tous les $A_\alpha \geqslant 0$ et sont indépendants du temps. Soit l'algorithme de décomposition

$$\begin{aligned} &\frac{\varphi^{j+1/n} - \varphi^j}{\tau} + A_1 \varphi^{j+1/n} = 0, \\ &\cdots\cdots\cdots\cdots\cdots \\ &\frac{\varphi^{j+1} - \varphi^{j+(n-1)/n}}{\tau} + A_n \varphi^{j+1} = 0. \end{aligned} \tag{4.38}$$

Montrons sa stabilité absolue. En effet, on multiplie

$$\frac{\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n}}{\tau} + A_\alpha \varphi^{j+\alpha/n} = 0 \tag{4.39}$$

scalairement par $\varphi^{j+\alpha/n}$, il vient

$$(\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n},\ \varphi^{j+\alpha/n}) + \tau (A_\alpha \varphi^{j+\alpha/n},\ \varphi^{j+\alpha/n}) = 0.$$

Comme A_α sont semi-définis positifs,

$$(\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n},\ \varphi^{j+\alpha/n}) \leqslant 0$$

ou

$$(\varphi^{j+\alpha/n},\ \varphi^{j+\alpha/n}) \leqslant (\varphi^{j+\alpha/n},\ \varphi^{j+(\alpha-1)/n}).$$

Etant donné

$$(\varphi^{j+\alpha/n},\ \varphi^{j+(\alpha-1)/n}) \leqslant \frac{1}{2} [(\varphi^{j+\alpha/n},\ \varphi^{j+\alpha/n}) + (\varphi^{j+(\alpha-1)/n},\ \varphi^{j+(\alpha-1)/n})],$$

on a

$$\| \varphi^{j+\alpha/n} \|^2 \leqslant \| \varphi^{j+(\alpha-1)/n} \|^2, \qquad \alpha = 1,\ 2,\ \ldots,\ n.$$

On aboutit par cette inégalité de récurrence à

$$\| \varphi^{j+1} \| \leqslant \| \varphi^j \|, \tag{4.40}$$

i.e. il y a stabilité numérique absolue sous les hypothèses faites.

Le système (4.38) approche le problème proposé à l'ordre 1 en τ, propriété vérifiée facilement.

Soit le problème non homogène

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + A\varphi = f \quad \text{dans} \quad D \times D_t, \\ &\varphi = g \quad \text{dans} \quad D \quad \text{pour} \quad t = 0 \end{aligned} \tag{4.41}$$

et le schéma de décomposition correspondant

$$\begin{aligned} &\frac{\varphi^{j+1/n} - \varphi^j}{\tau} + A_1 \varphi^{j+1/n} = 0, \\ &\cdots\cdots\cdots\cdots\cdots \\ &\frac{\varphi^{j+1} - \varphi^{j+(n-1)/n}}{\tau} + A_n \varphi^{j+1} = f^j. \end{aligned} \tag{4.42}$$

Ce schéma approche l'équation non homogène considérée à des quantités d'ordre 1 près en τ.

Afin de démontrer la stabilité, on multiplie chaque équation scalairement par $\varphi^{j+1/n}, \ldots, \varphi^{j+1}$ respectivement. On a comme plus haut

$$\| \varphi^{j+\alpha/n} \| \leqslant \| \varphi^{j+(\alpha-1)/n} \| \quad (\alpha = 1, 2, \ldots, n-1). \tag{4.43}$$

Arrêtons-nous sur la dernière équation (4.42). On a

$$(\varphi^{j+1}, \varphi^{j+1}) = (\varphi^{j+(n-1)/n}, \varphi^{j+1}) - \tau (A_n \varphi^{j+1}, \varphi^{j+1}) + \tau (f^j, \varphi^{j+1}).$$

Vu que $A_n \geqslant 0$,

$$(\varphi^{j+1}, \varphi^{j+1}) \leqslant (\varphi^{j+(n-1)/n}, \varphi^{j+1}) + \tau (f^j, \varphi^{j+1}).$$

L'inégalité de Cauchy-Bouniakovski entraîne

$$|(\varphi^{j+(n-1)/n}, \varphi^{j+1})| \leqslant \| \varphi^{j+(n-1)/n} \| \| \varphi^{j+1} \|,$$
$$|(f^j, \varphi^{j+1})| \leqslant \| f^j \| \| \varphi^{j+1} \|.$$

Donc

$$\| \varphi^{j+1} \|^2 \leqslant \| \varphi^{j+(n-1)/n} \| \| \varphi^{j+1} \| + \tau \| f^j \| \| \varphi^{j+1} \|.$$

On obtient en simplifiant par $\| \varphi^{j+1} \|$:

$$\| \varphi^{j+1} \| \leqslant \| \varphi^{j+(n-1)/n} \| + \tau \| f^j \|$$

et, par élimination des solutions munies d'indices non entiers,

$$\| \varphi^{j+1} \| \leqslant \| \varphi^j \| + \tau \| f^j \|. \tag{4.44}$$

Puisque

$$\| \varphi^0 \| = \| g \|,$$

on a en éliminant les valeurs intermédiaires

$$\| \varphi^{j+1} \| \leqslant \| g \| + \tau j \| f \|, \tag{4.45}$$

où

$$\| f \| = \max_j \| f^j \|,$$

d'où la stabilité absolue du schéma aux différences.

Cet algorithme se transpose à A dépendant du temps, auquel cas on remplace cet opérateur par son approximation aux différences convenable associée à chaque intervalle $t_j \leqslant t \leqslant t_{j+1}$.

5.4.4. Décomposition des problèmes quasi linéaires

Soit un problème d'évolution à A dépendant du temps et de la solution, à savoir

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + A(t, \varphi)\varphi = 0 \quad \text{dans} \quad D \times D_t, \\ &\varphi = g \quad \text{dans} \quad D \quad \text{pour} \quad t = 0. \end{aligned} \tag{4.46}$$

L'opérateur $A(t, \varphi)$ est supposé non négatif tel que

$$A(t, \varphi) = \sum_{\alpha=1}^{n} A_\alpha(t, \varphi), \tag{4.47}$$

$A_\alpha(t, \varphi) \geqslant 0$ et est suffisamment régulier. On suppose de plus que la solution φ est elle aussi une fonction du temps suffisamment régulière. Considérons sur l'intervalle $t_{j-1} \leqslant t \leqslant t_{j+1}$ le schéma de décomposition

$$\begin{aligned}
&\frac{\varphi^{j+1/n-1} - \varphi^{j-1}}{\tau} + A_1^j \frac{\varphi^{j+1/n-1} + \varphi^{j-1}}{2} = 0,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\frac{\varphi^{j} - \varphi^{j-1/n}}{\tau} + A_n^j \frac{\varphi^{j} + \varphi^{j-1/n}}{2} = 0,\\
&\frac{\varphi^{j+1/n} - \varphi^{j}}{\tau} + A_n^j \frac{\varphi^{j+1/n} + \varphi^{j}}{2} = 0,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\frac{\varphi^{j+1} - \varphi^{j+(n-1)/n}}{\tau} + A_1^j \frac{\varphi^{j+1} + \varphi^{j+(n-1)/n}}{2} = 0,
\end{aligned} \tag{4.48}$$

où

$$\begin{aligned}
&A^j = A(t_j, \tilde{\varphi}^j),\\
&\tilde{\varphi}^j = \varphi^{j-1} - \tau A(t_{j-1}, \varphi^{j-1})\varphi^{j-1},\\
&\tau = t_j - t_{j-1}.
\end{aligned} \tag{4.49}$$

On procède comme pour les opérateurs linéaires dépendant du temps seul et on démontre sans peine que sous la condition (4.49) le schéma (4.48) est d'ordre d'approximation 2 en τ et absolument stable. On traite de même les équations quasi linéaires non homogènes. Cette circonstance ouvre des horizons nouveaux à la méthode de splitting-up dans les problèmes quasi linéaires non stationnaires de très nombreux domaines des sciences de la nature : hydrodynamique, météorologie, océanographie.

5.5. Méthode de splitting-up (approche générale)

Le problème suivant se pose fréquemment en physique mathématique : décomposer les équations différentielles, intégrales ou intégro-différentielles en des équations plus simples qu'on met ensuite sous forme de différences finies par les techniques de ce chapitre. Il est intimement lié au problème d'approximation faible étudié par Samarski [III ; 12] [XV ; 5], Yanenko [III], Yanenko, Démidov [XV ; 14], Diakonov, Lébédev [XV ; 6] et développé par beaucoup d'auteurs. C'est également notre but à nous.

Soit un problème de la physique mathématique

$$\frac{\partial \varphi}{\partial t} + A\varphi = 0 \quad \text{dans} \quad D \times D_t,$$
$$\varphi = g \quad \text{dans} \quad D \quad \text{pour} \quad t = 0. \tag{5.1}$$

Faisons l'hypothèse

$$A = \sum_{\alpha=1}^{n} A_\alpha, \tag{5.2}$$

et $A_\alpha \geqslant 0$. La solution φ et la fonction f sont supposées suffisamment régulières. On représente (5.1) sur chaque intervalle $\theta_j = \{t_j \leqslant t \leqslant t_{j+1}\}$ par le problème

$$\frac{\partial \varphi_\alpha}{\partial t} + A_\alpha \varphi_\alpha = 0 \quad \text{dans} \quad D \times \theta_j,$$
$$\varphi_\alpha^j = \varphi_{\alpha-1}^{j+1} \quad \text{dans} \quad D \quad \text{pour} \quad t = t_j, \tag{5.3}$$
$$\alpha = 1, 2, \ldots, n,$$

où

$$\varphi_0^{j+1} = \varphi^j, \qquad \varphi_n^{j+1} = \varphi^{j+1}. \tag{5.4}$$

Il est déjà connu qu'en discrétisant chaque équation à l'aide du schéma de Crank-Nicholson, on obtient le système

$$\frac{\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n}}{\tau} + A_\alpha \frac{\varphi^{j+\alpha/n} + \varphi^{j+(\alpha-1)/n}}{2} = 0,$$
$$\alpha = 1, 2, \ldots, n, \tag{5.5}$$

où

$$\varphi^{j+\alpha/n} = \varphi_\alpha^{j+1}, \qquad \varphi^{j+1} = \varphi_n^{j+1}. \tag{5.6}$$

Supposons que chaque A_α s'écrit comme

$$A_\alpha = \sum_{\beta=1}^{m_\alpha} A_{\alpha\beta}, \tag{5.7}$$

$A_{\alpha\beta} \geqslant 0$. Une question se pose : ne vaut-il pas mieux représenter dès le début A par $A_{\alpha\beta}$ au lieu de décomposer A en A_α, puis A_α en $A_{\alpha\beta}$? Malgré l'équivalence apparente des deux procédés, on préfère dans nombre de cas de remplacer un gros problème par une succession de problèmes d'importance moyenne qu'on réduit ensuite de façon indépendante et aisée à des problèmes simples.

Prenons le système (5.3) et décomposons-le, compte tenu de (5.7),

$$\frac{\varphi_\alpha^{j+\beta/m_\alpha} - \varphi_\alpha^{j+(\beta-1)/m_\alpha}}{\tau} + A_{\alpha\beta} \frac{\varphi_\alpha^{j+\beta/m_\alpha} + \varphi_\alpha^{j+(\beta-1)/m_\alpha}}{2} = 0, \tag{5.8}$$
$$\alpha = 1, 2, \ldots, n; \quad \beta = 1, 2, \ldots, m_\alpha,$$

avec

$$\varphi_1^j = \varphi^j; \quad \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1} \quad (\alpha > 1); \quad \varphi_n^{j+1} = \varphi^{j+1}.$$

On constate facilement que (5.8) approche le problème proposé (5.1) à des quantités du second ordre près en τ s'il y a commutativité des opérateurs $A_{\alpha\beta}$. En effet, si l'on change le numérotage, compte tenu de (5.2) et (5.7):

$$A = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{m_\alpha} A_{\alpha\beta} = \sum_{\gamma=1}^{p} A_\gamma,$$

on a le problème

$$\frac{\varphi^{j+\gamma/p} - \varphi^{j+(\gamma-1)/p}}{\tau} + A_\gamma \frac{\varphi^{j+\gamma/p} + \varphi^{j+(\gamma-1)/p}}{2} = 0, \qquad (5.9)$$

$$\gamma = 1, 2, \ldots, p,$$

qui approche, on le sait dès le n° 4.4, le problème (5.1) à l'ordre 2 en τ. Le résultat subsiste pour $A_{\alpha\beta}$ dépendant du temps, auquel cas on approche sur chaque $t_j \leqslant t \leqslant t_{j+1}$ les opérateurs $A_{\alpha\beta} = \Lambda_{\alpha\beta}^j$ à l'ordre 2 en τ. Si $\Lambda_{\alpha\beta}^j$ ne commutent pas, la procédure 2-cyclique du n° 4.4 conduit à un schéma exact à l'ordre 2 sur chaque $t_{j-1} \leqslant t \leqslant t_{j+1}$.

Faisons le bilan. Supposons qu'on ramène un problème d'évolution (5.1), $A_\alpha \geqslant 0$, à des problèmes particuliers (5.3) et qu'on assimile chacun de ceux-ci à un ensemble de problèmes élémentaires. Si l'un au moins de ces problèmes élémentaires se réduit à des schémas aux différences exacts à l'ordre 1, alors (5.1) est approché à l'ordre 1 en τ. Si chaque problème élémentaire est approché par un schéma exact à l'ordre 2, on est conduit par un algorithme 2-cyclique en α et β à une approximation d'ordre 2 en τ. On note qu'avec $A_{\alpha\beta}$ non commutatifs et en l'absence de la procédure 2-cyclique, le problème (5.1) n'est approché qu'à l'ordre 1 près.

Dans le cas non commutatif, on a au départ le problème

$$\frac{\partial \varphi}{\partial t} + \sum_{\alpha=1}^{n} A_\alpha \varphi = 0 \quad \text{dans} \quad D \times \theta_j,$$
$$\varphi = \varphi^j \text{ dans } D \text{ pour } t = t_j. \qquad (5.10)$$

On lui associe le système

$$\frac{\partial \varphi_\alpha}{\partial t} + A_\alpha \varphi_\alpha = 0, \quad \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1}, \quad \alpha = 1, 2, \ldots, n. \qquad (5.11)$$

Soit $A_\alpha = \sum_{\beta=1}^{m_\alpha} A_{\alpha\beta}$. On applique à chaque problème (5.11) la méthode 2-cyclique:

$$\frac{\varphi_\alpha^{j+\frac{\beta}{2m_\alpha}} - \varphi_\alpha^{j+\frac{\beta-1}{2m_\alpha}}}{\tau/2} + A_{\alpha\beta}\frac{\varphi^{j+\frac{\beta}{2m_\alpha}} + \varphi^{j+\frac{\beta-1}{2m_\alpha}}}{2} = 0,$$

$$\beta = 1, 2, \ldots, m_\alpha,$$

$$\frac{\varphi_\alpha^{j+\frac{\beta}{2m_\alpha}} - \varphi_\alpha^{j+\frac{\beta-1}{2m_\alpha}}}{\tau/2} + A_{\alpha,\, 2m_\alpha+1-\beta}\frac{\varphi^{j+\frac{\beta}{2m_\alpha}} + \varphi^{j+\frac{\beta-1}{2m_\alpha}}}{2} = 0, \tag{5.12}$$

$$\beta = m_\alpha + 1, \quad m_\alpha + 2, \ldots, 2m_\alpha.$$

Les conditions initiales s'écrivent respectivement

$$\varphi_1^j = \varphi^j, \qquad \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1} \qquad (\alpha = 2, \ldots, n). \tag{5.13}$$

On voit sans peine que le problème (5.12) approche tout (5.11) sur $t_j \leqslant t \leqslant t_{j+1}$ à τ^2 près.

Pour que tout l'algorithme résout (5.1) à la même précision, on doit alterner les cycles principaux. Ainsi, on a sur $t_{j-1} \leqslant t \leqslant t_j$ au lieu de (5.11):

$$\frac{\partial \varphi_\alpha}{\partial t} + A_\alpha \varphi_\alpha = 0, \qquad \alpha = 1, 2, \ldots, n,$$
$$\varphi_1^{j-1} = \varphi^{j-1}, \qquad \varphi_\alpha^{j-1} = \varphi_{\alpha-1}^j \qquad (\alpha > 1), \qquad \varphi^j = \varphi_n^j, \tag{5.14}$$

et sur $t_j \leqslant t \leqslant t_{j+1}$

$$\frac{\partial \varphi_\alpha}{\partial t} + A_{n-\alpha+1}\varphi_\alpha = 0, \qquad \alpha = 1, 2, \ldots, n,$$
$$\varphi_1^j = \varphi^j, \quad \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1} \quad (\alpha > 1), \quad \varphi^{j+1} = \varphi_n^{j+1}. \tag{5.15}$$

On suppose de plus que chaque problème (5.14) et (5.15) est abordé par une technique 2-cyclique (5.12). On note qu'avec $A_{\alpha\beta} \geqslant 0$, la méthode de splitting-up est absolument stable.

Etablissons enfin un schéma général de décomposition pour l'équation non homogène

$$\frac{\partial \varphi}{\partial t} + \sum_{\alpha=1}^{n} A_\alpha \varphi = f,$$
$$\varphi = g \quad \text{pour} \quad t = 0 \tag{5.16}$$

sur l'intervalle $t_{j-1} \leqslant t \leqslant t_{j+1}$ dans le cadre de la méthode 2-cyclique. Considérons des schémas d'approximation faible sous forme différentielle.

On pose sur $t_{j-1} \leqslant t \leqslant t_j$:

$$\begin{aligned} &\frac{\partial \varphi_\alpha}{\partial t} + A_\alpha \varphi_\alpha = 0, \qquad \alpha = 1, 2, \ldots, n-1, \\ &\frac{\partial \varphi_n}{\partial t} + A_n \varphi_n = f + \frac{\tau}{2} A_n f, \end{aligned} \tag{5.17}$$

et sur $t_j \leqslant t \leqslant t_{j+1}$

$$\begin{aligned} &\frac{\partial \varphi_{n+1}}{\partial t} + A_n \varphi_{n+1} = f - \frac{\tau}{2} A_n f, \\ &\frac{\partial \varphi_{n+\alpha}}{\partial t} + A_{n-\alpha+1} \varphi_{n+\alpha} = 0, \qquad \alpha = 2, 3, \ldots, n, \end{aligned} \tag{5.18}$$

sous les conditions

$$\varphi_1 (t_{j-1}) = \varphi (t_{j-1}), \quad \varphi_{\alpha+1} (t_{j-1}) = \\ = \varphi_\alpha (t_j), \quad \alpha = 1, 2, \ldots, n, \tag{5.19}$$

et

$$\varphi_{\alpha+1} (t_j) = \varphi_\alpha (t_{j+1}), \qquad \alpha = n+1, n+2, \ldots, 2n. \tag{5.20}$$

Si l'on résout (5.17)-(5.19) sur $t_{j-1} \leqslant t \leqslant t_{j+1}$ moyennant les schémas de Crank-Nicholson et si l'on pose $f = f^j$, on obtient (4.35) pour (5.18).

A part le système (5.17), (5.18), on considère:

$$\begin{aligned} &\frac{\partial \varphi_1}{\partial t} + A_1 \varphi_1 = 0, \\ &\ldots\ldots\ldots \\ &\frac{\partial \varphi_n}{\partial t} + A_n \varphi_n = 0 \end{aligned} \tag{5.21}$$

sur $t_{j-1} \leqslant t \leqslant t_j$;

$$\frac{\partial \varphi_{n+1}}{\partial t} = f \tag{5.22}$$

sur $t_{j-1} \leqslant t \leqslant t_{j+1}$;

$$\begin{aligned} &\frac{\partial \varphi_{n+2}}{\partial t} + A_n \varphi_{n+2} = 0, \\ &\ldots\ldots\ldots\ldots \\ &\frac{\partial \varphi_{2n+1}}{\partial t} + A_1 \varphi_{2n+1} = 0 \end{aligned} \tag{5.23}$$

sur $t_j \leqslant t \leqslant t_{j+1}$. Les conditions initiales sont respectivement

$$\varphi_1 (t_{j-1}) = \varphi (t_{j-1}), \quad \varphi_\alpha (t_{j-1}) = \varphi_{\alpha-1} (t_j) \qquad (\alpha = 2, 3, \ldots, n), \tag{5.24}$$

$$\varphi_{n+1} (t_{j-1}) = \varphi_n (t_j) \tag{5.25}$$

et

$$\varphi_\alpha(t_j) = \varphi_{\alpha-1}(t_{j+1}), \qquad \alpha = n+2,\ n+3,\ \ldots,\ 2n+1. \tag{5.26}$$

La propriété d'approximation et la stabilité des schémas obtenus en garantissent la convergence (voir n° 1.4).

5.6. Méthodes de résolution des équations hyperboliques

Une vaste classe de problèmes de la physique mathématique relève des équations hyperboliques dont la résolution numérique date de l'ouvrage fondamental de Courant, Friedrichs, Lewy [VII]. De nombreux mathématiciens s'y sont intéressés intensément (citons entre autres Ladyjenskaïa [II; 5], Godounov [II], Samarski [III; 12], Konovalov [XV; 8]). On connaît plusieurs méthodes efficaces pour les problèmes hyperboliques attachés à la théorie des vibrations et à la théorie de l'élasticité pour le cas des domaines multidimensionnels. S'agissant des opérateurs, solutions et données réguliers, ces méthodes procèdent d'algorithmes spéciaux ci-dessous.

5.6.1. Méthode de stabilisation

Soit le problème

$$\begin{aligned} &\frac{\partial^2\varphi}{\partial t^2} + A\varphi = f \quad \text{dans} \quad D \times D_t, \\ &\varphi = p, \quad \frac{\partial\varphi}{\partial t} = q \quad \text{dans} \quad D \quad \text{pour} \quad t = 0, \end{aligned} \tag{6.1}$$

où A est supposé être un opérateur aux différences indépendant du temps et p et q sont telles qu'il y ait régularité suffisante de la solution. Faisons l'hypothèse

$$(A\varphi,\ \varphi) \geqslant \gamma^2(\varphi,\ \varphi), \tag{6.2}$$

i.e. A est défini positif. On rappelle qu'on a pour A définis positifs

$$\gamma^2 = \alpha\left(\frac{A^*+A}{2}\right),$$

avec α la plus petite valeur propre de l'opérateur $\frac{A+A^*}{2}$.

Considérons le problème aux différences qui approche (6.1)

$$\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + A\varphi^j = f^j. \tag{6.3}$$

On montre aisément qu'étant donnée la régularité de la solution, le schéma (6.3) approche l'équation de (6.1) à l'ordre 2 en τ. Adjoignons à (6.3) les conditions initiales. On veut conserver l'ordre d'ap-

proximation et on prend donc les conditions

$$\varphi^0 = p, \quad \varphi^1 = \left(E - \frac{\tau^2}{2} A\right) p + \tau q + \frac{\tau^2}{2} f^0. \tag{6.4}$$

La dernière relation s'obtient en développant la solution de (6.1) en série de Taylor autour de $t = 0$ et en éliminant les dérivées à l'aide de l'équation et des conditions initiales données du problème (6.1).

Ainsi, le problème (6.3), (6.4) se trouve posé. Quelle est la stabilité numérique du schéma (6.3)? On l'établit en termes de spectre.

Soient u_n, u_n^* des fonctions propres et $\lambda_n > 0$ les valeurs propres associées des problèmes spectraux

$$Au = \lambda u, \quad A^* u^* = \lambda u^*. \tag{6.5}$$

On suppose que $\{u_n\}$ forme une base et on cherche une solution de la forme

$$\varphi^j = \sum_n \varphi_n^j u_n, \tag{6.6}$$

où

$$\varphi_n^j = (\varphi^j, u_n^*).$$

Portons la série de Fourier (6.6) dans (6.3) et multiplions par u_n^*, il vient pour les coefficients de Fourier φ_n^j

$$\frac{\varphi_n^{j+1} - 2\varphi_n^j + \varphi_n^{j-1}}{\tau^2} + \lambda_n \varphi_n^j = f_n^j. \tag{6.7}$$

On cherche la solution générale de l'équation homogène associée sous forme de fonction puissance

$$\varphi_n^j = \eta_n^j \tag{6.8}$$

(j du premier membre est un indice et j à droite un exposant).

On porte (6.8) dans (6.7) et on pose $f_n^j = 0$. On aboutit à l'équation caractéristique pour η_n

$$\eta_n^2 - 2\left(1 - \frac{\tau^2 \lambda_n}{2}\right) \eta_n + 1 = 0. \tag{6.9}$$

On vérifie de suite que si

$$\left|1 - \frac{\tau^2 \lambda_n}{2}\right| \leqslant 1, \tag{6.10}$$

alors l'équation (6.9) a ses racines complexes conjuguées de module $= 1$, i.e.

$$|\eta_n| = 1. \tag{6.11}$$

On tire de (6.10)

$$\tau^2 \leqslant \frac{4}{\lambda_n}, \qquad n = 1, 2, \ldots. \tag{6.12}$$

Evidemment, (6.12) a lieu pour tout λ_n si

$$\tau \leqslant \frac{2}{\sqrt{\beta(A)}}, \tag{6.13}$$

$\beta(A)$ étant la borne supérieure du spectre de l'opérateur A. S'agissant des opérateurs symétriques on a $\beta(A) = \| A \|$, donc

$$\tau \leqslant \frac{2}{\sqrt{\|A^2\|}}. \tag{6.14}$$

On passe aux schémas aux différences implicites

$$\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + A\frac{\varphi^{j+1} + \varphi^{j-1}}{2} = f^j. \tag{6.15}$$

Ce schéma est exact à l'ordre 2 en τ et le problème (6.15), (6.4) approche (6.1) à l'ordre 2. L'équation caractéristique correspondante s'écrit

$$\eta_n^2 - \frac{2}{1 + \frac{\tau^2\lambda_n}{2}}\eta_n + 1 = 0, \tag{6.16}$$

par conséquent,

$$\eta_n = \frac{1}{1 + \frac{\tau^2\lambda_n}{2}} \pm \sqrt{\left(\frac{1}{1 + \frac{\tau^2\lambda_n}{2}}\right)^2 - 1}\,. \tag{6.17}$$

Ainsi, on a

$$|\eta_n| = 1 \tag{6.18}$$

quels que soient τ et n. Le schéma (6.15) est absolument stable (voir Richtmyer, Morton [III]).

Plaçons-nous au cas

$$A = \sum_{\alpha=1}^{n} A_\alpha, \tag{6.19}$$

et tous les $A_\alpha \geqslant 0$. On utilise, pour approcher (6.1), l'équation aux différences de la forme

$$B\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + A\varphi^j = f^j, \tag{6.20}$$

où

$$B = \prod_{\alpha=1}^{n}\left(E + \frac{\tau^2}{2}A_\alpha\right). \tag{6.21}$$

Il résulte de (6.20) et (6.21) que l'équation (6.20) approche (6.1) à des quantités du second ordre près en τ. Comme elle se ramène à

$$\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + B^{-1}A\varphi^j = B^{-1}f^j, \tag{6.22}$$

la théorie des séries de Fourier entraîne la condition nécessaire de stabilité du schéma (6.20), (6.21)

$$\tau \leqslant \frac{2}{\sqrt{\beta (B^{-1}A)}}. \tag{6.23}$$

Ainsi, choisir un paramètre τ vérifiant la condition de stabilité équivaut à calculer la plus grande valeur propre du problème

$$Au = \lambda Bu, \tag{6.24}$$

la matrice $B^{-1}A$ étant supposée à valeurs propres positives. On résout ce problème par la méthode itérative de Ljusternik.

Le schéma aux différences associé à (6.20) est mis en œuvre de façon suivante

$$\begin{aligned}
&\left(E + \frac{\tau^2}{2} A_1\right) \xi^{j+1/n} = -A\varphi^j + f^j, \\
&\left(E + \frac{\tau^2}{2} A_2\right) \xi^{j+2/n} = \xi^{j+1/n}, \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
&\left(E + \frac{\tau^2}{2} A_n\right) \xi^{j+1} = \xi^{j+(n-1)/n}, \\
&\varphi^{j+1} = 2\varphi^j - \varphi^{j-1} + \tau^2 \xi^{j+1}.
\end{aligned} \tag{6.25}$$

On utilise successivement $j = 2, 3, \ldots$ et les données initiales (6.4). Le schéma (6.20) constitue un schéma de décomposition.

5.6.2. Réduction de l'équation des vibrations à un problème d'évolution

La construction de schémas aux différences absolument stables pour les équations hyperboliques qui approchent le problème proposé à l'ordre 2 et soient efficaces en calcul automatique a exigé des méthodes de décomposition spéciales analogues à celles attachées aux problèmes d'évolution.

L'idée de base de la réduction formelle d'un problème hyperbolique à un problème d'évolution est illustrée par le problème simple des membranes vibrantes avec les conditions aux limites périodiques par rapport au carré $D = \{0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1\}$:

$$\begin{aligned}
&\frac{\partial^2 \varphi}{\partial t^2} = \frac{\partial}{\partial x} a^2 \frac{\partial \varphi}{\partial x} + \frac{\partial}{\partial y} a^2 \frac{\partial \varphi}{\partial y} \quad \text{dans} \quad D \times D_t, \\
&\varphi = p, \qquad \frac{\partial \varphi}{\partial t} = q \ \text{dans} \ D \ \text{pour} \ t = 0.
\end{aligned} \tag{6.26}$$

Ici $a^2 = a^2(x, y)$ est le carré de la vitesse de propagation des perturbations, $p = p(x, y)$, $q = q(x, y)$ sont des fonctions données. La solution périodique φ est suffisamment régulière pour qu'on puisse

effectuer des transformations nécessaires et former des schémas aux différences exacts à l'ordre 2 en x, y, t.

L'équation de (6.26) se met sous la forme

$$\begin{aligned} &\frac{\partial u}{\partial t} - a\frac{\partial \varphi}{\partial x} = 0, \\ &\frac{\partial v}{\partial t} - a\frac{\partial \varphi}{\partial y} = 0 \qquad \text{dans} \quad D \times D_t, \\ &\frac{\partial \varphi}{\partial t} - \left(\frac{\partial au}{\partial x} + \frac{\partial av}{\partial y}\right) = 0. \end{aligned} \tag{6.27}$$

Les données initiales sont

$$u = u^0(x, y), \quad v = v^0(x, y), \quad \varphi = p(x, y) \quad \text{pour} \quad t = 0. \tag{6.28}$$

Le choix de u^0 et v^0 peut être plus ou moins arbitraire à condition d'avoir la relation

$$\frac{\partial au^0}{\partial x} + \frac{\partial av^0}{\partial y} = q(x, y). \tag{6.29}$$

Introduisons une matrice A et un vecteur φ définis par

$$A = \begin{Vmatrix} 0 & 0 & -a\frac{\partial}{\partial x} \\ 0 & 0 & -a\frac{\partial}{\partial y} \\ -\frac{\partial}{\partial x}a & -\frac{\partial}{\partial y}a & 0 \end{Vmatrix}, \qquad \varphi = \begin{Vmatrix} u \\ v \\ \varphi \end{Vmatrix},$$

auquel cas le système (6.27) et les données initiales (6.28) s'écrivent

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + A\varphi = 0 \quad \text{dans} \quad D \times D_t, \\ &\varphi = \varphi^0 \quad \text{dans} \quad D \quad \text{pour} \quad t = 0, \end{aligned} \tag{6.30}$$

φ^0 ayant pour composantes u^0, v^0 et p.

Pour établir la définitivité de l'opérateur A on forme la fonctionnelle

$$(A\varphi, \varphi) = -\int_D \left[\frac{\partial}{\partial x}(au\varphi) + \frac{\partial}{\partial y}(av\varphi)\right] dD = -\int_S au_n\varphi \, dS, \tag{6.31}$$

avec u_n la composante normale à S du vecteur $\mathbf{u} = u\mathbf{i} + v\mathbf{j}$. Etant donnée la périodicité de a, φ (et des dérivées de la solution), les valeurs de u_n aux points des côtés opposés de D et symétriques par rapport à son centre sont évidemment de même module et de signes contraires. L'intégrale de surface de (6.31) s'annule donc et on est conduit à la condition

$$(A\varphi, \varphi) = 0. \tag{6.32}$$

On note que si l'on impose, au lieu de la périodicité, la fixation rigide de la membrane ($\varphi = 0$), on a la condition (6.32) (cela découle de (6.31)). Cette remarque s'applique évidemment à D de forme quelconque. La condition (6.32) garantit l'unicité de la solution. En effet, multiplions (6.30) scalairement par φ et utilisons (6.32), il vient

$$\frac{d}{dt} \| \varphi \|^2 = 0, \tag{6.33}$$

où

$$\| \varphi \| = \left[\int_D (u^2 + v^2 + \varphi^2)\, dD \right]^{1/2}.$$

On suppose

$$u^0 = 0, \quad v^0 = 0, \quad \varphi^0 = 0,$$

auquel cas

$$\| \varphi^0 \| = 0. \tag{6.34}$$

Résolvons (6.33) sous cette condition :

$$\| \varphi \| = \| \varphi^0 \| = 0,$$

i.e. on a à tout instant

$$u = 0, \quad v = 0, \quad \varphi = 0,$$

ce qui démontre l'unicité.

Formulons la méthode de décomposition pour le problème (6.30). On introduit les matrices

$$A_1 = \begin{Vmatrix} 0 & 0 & -a\frac{\partial}{\partial x} \\ 0 & 0 & 0 \\ -\frac{\partial}{\partial x} a & 0 & 0 \end{Vmatrix}, \qquad A_2 = \begin{Vmatrix} 0 & 0 & 0 \\ 0 & 0 & -a\frac{\partial}{\partial y} \\ 0 & -\frac{\partial}{\partial y} a & 0 \end{Vmatrix}.$$

Evidemment,

$$A = A_1 + A_2. \tag{6.35}$$

On montre comme plus haut que

$$(A_1\varphi, \varphi) = 0, \quad (A_2\varphi, \varphi) = 0. \tag{6.36}$$

Autrement dit, on aborde le problème proposé sur chaque intervalle $t_j \leqslant t \leqslant t_{j+1}$ par n'importe quel procédé du numéro 4.3 (stabilisation, prédiction-correction, splitting-up). Si l'équation des vibrations est à plusieurs variables, on opte pour la dernière méthode qui conduit à des schémas absolument stables exacts à l'ordre 2 sous des conditions de la forme (6.36).

Effectuons la réduction du problème (6.30) sur chaque $t_j \leqslant t \leqslant t_{j+1}$ à l'aide de cette technique. On a

$$\begin{aligned} &\frac{\varphi^{j+1/2}-\varphi^j}{\tau}+A_1\frac{\varphi^{j+1/2}+\varphi^j}{2}=0,\\ &\frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau}+A_2\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}=0, \end{aligned} \tag{6.37}$$

ou scalairement

$$\begin{aligned} &\frac{u^{j+1/2}-u^j}{\tau}=a\frac{\partial}{\partial x}\left(\frac{\varphi^{j+1/2}+\varphi^j}{2}\right),\\ &\frac{v^{j+1/2}-v^j}{\tau}=0,\\ &\frac{\varphi^{j+1/2}-\varphi^j}{\tau}=\frac{\partial}{\partial x}\left(a\frac{u^{j+1/2}+u^j}{2}\right) \end{aligned} \tag{6.38}$$

et

$$\begin{aligned} &\frac{u^{j+1}-u^{j+1/2}}{\tau}=0,\\ &\frac{v^{j+1}-v^{j+1/2}}{\tau}=a\frac{\partial}{\partial y}\left(\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}\right),\\ &\frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau}=\frac{\partial}{\partial y}\left(a\frac{v^{j+1}+v^{j+1/2}}{2}\right). \end{aligned} \tag{6.39}$$

Vu que $v^{j+1/2}=v^j$ et $u^{j+1/2}=u^{j+1}$, les systèmes (6.38) se simplifient comme suit :

$$\begin{aligned} &\frac{u^{j+1}-u^j}{\tau}=a\frac{\partial}{\partial x}\left(\frac{\varphi^{j+1/2}+\varphi^j}{2}\right),\\ &\frac{\varphi^{j+1/2}-\varphi^j}{\tau}=\frac{\partial}{\partial x}\left(a\frac{u^{j+1}+u^j}{2}\right), \end{aligned} \tag{6.40}$$

$$\begin{aligned} &\frac{v^{j+1}-v^j}{\tau}=a\frac{\partial}{\partial y}\left(\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}\right),\\ &\frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau}=\frac{\partial}{\partial y}\left(a\frac{v^{j+1}+v^j}{2}\right). \end{aligned} \tag{6.41}$$

On trouve u^{j+1}, $\varphi^{j+1/2}$ à partir de (6.40) et v^{j+1}, φ^{j+1} à partir de (6.41).

On discrétise en x et y de façon à obtenir pour u, v et φ des schémas absolument stables d'ordre d'approximation 2 et à rester dans les conditions (6.36) pour A_1 et A_2 sous forme de différences

finies. On pose

$$\frac{u_{k,\,l}^{j+1}-u_{k,\,l}^{j}}{\tau}=\frac{a_{k,\,l}}{h}\left[\left(\frac{\varphi^{j+1/2}+\varphi^{j}}{2}\right)_{k,\,l}-\left(\frac{\varphi^{j+1/2}+\varphi^{j}}{2}\right)_{k-1,\,l}\right], \tag{6.42}$$

$$\frac{\varphi_{k,\,l}^{j+1/2}-\varphi_{k,\,l}^{j}}{\tau}=\frac{1}{h}\left[a_{k+1,\,l}\left(\frac{u^{j+1}+u^{j}}{2}\right)_{k+1,\,l}-a_{k,\,l}\left(\frac{u^{j+1}+u^{j}}{2}\right)_{k,\,l}\right].$$

Ainsi, la première équation utilise une différence « arrière » et la seconde une différence « avant ». On a de même pour (6.41)

$$\frac{v_{k,\,l}^{j+1}-v_{k,\,l}^{j}}{\tau}=\frac{a_{k,\,l}}{h}\left[\left(\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}\right)_{k,\,l}-\left(\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}\right)_{k,\,l-1}\right], \tag{6.43}$$

$$\frac{\varphi_{k,\,l}^{j+1}-\varphi_{k,\,l}^{j+1/2}}{\tau}=\frac{1}{h}\left[a_{k,\,l+1}\left(\frac{v^{j+1}+v^{j}}{2}\right)_{k,\,l+1}-a_{k,\,l}\left(\frac{v^{j+1}-v^{j}}{2}\right)_{k,\,l}\right].$$

On écrit (6.42), (6.43) pour $k = 1, 2, \ldots, N-1$, $l = 1, 2, \ldots, N-1$.

Considérons le cas avec la condition à la frontière de D

$$\varphi = 0 \text{ sur } \partial D \times D_t. \tag{6.44}$$

On a, en projetant (6.44) sur le réseau $\partial D_h \times D_\tau$:

$$\varphi_{0,\,l}^{j}=\varphi_{N,\,l}^{j}=0 \quad \text{et} \quad \varphi_{k,\,0}^{j}=\varphi_{k,\,N}^{j}=0, \tag{6.45}$$

j pouvant être entier ou fractionnaire. On note que connaissant $\varphi_{k,\,l}^{j}$, $\varphi_{k,\,l}^{j+1/2}$, $\varphi_{k,\,l}^{j+1}$ on définit $u_{k,\,l}^{j+1}$ et $v_{k,\,l}^{j+1}$ en tous les nœuds frontières à l'aide de (6.42), (6.43) et (6.45).

Les équations (6.42), (6.43) se ramènent par élimination des inconnues $u_{k,\,l}^{j+1}$, $v_{k,\,l}^{j+1}$ à des équations aux différences en $\varphi_{k,\,l}$. On introduit à cet effet les quantités auxiliaires

$$\varphi_{k,\,l}^{j+1/4}=\frac{1}{2}\left(\varphi_{k,\,l}^{j+1/2}+\varphi_{k,\,l}^{j}\right), \qquad \varphi_{k,\,l}^{j+3/4}=\frac{1}{2}\left(\varphi_{k,\,l}^{j+1}+\varphi_{k,\,l}^{j+1/2}\right), \tag{6.46}$$

auquel cas (6.42), (6.43) se récrivent

$$\mu_{k+1,\,l}^{2}\left(\varphi_{k+1,\,l}^{j+1/4}-\varphi_{k,\,l}^{j+1/4}\right)-\mu_{k,\,l}^{2}\left(\varphi_{k,\,l}^{j+1/4}-\varphi_{k-1,\,l}^{j+1/4}\right)-\varphi_{k,\,l}^{j+1/4}=-f_{k,\,l}^{j+1/4}, \tag{6.47}$$

$$\mu_{k,\,l+1}^{2}\left(\varphi_{k,\,l+1}^{j+3/4}-\varphi_{k,\,l}^{j+3/4}\right)-\mu_{k,\,l}^{2}\left(\varphi_{k,\,l}^{j+3/4}-\varphi_{k,\,l-1}^{j+3/4}\right)-\varphi_{k,\,l}^{j+3/4}=-f_{k,\,l}^{j+3/4},$$

où

$$\mu_{k,\,l}=\frac{\tau a_{k,\,l}}{2h},$$

$$\begin{aligned} f_{k,\,l}^{j+1/4} &= 2\varphi_{k,\,l}^{j}+2\,(\mu_{k+1,\,l}\,u_{k+1,\,l}^{j}-\mu_{k,\,l}\,u_{k,\,l}^{j}),\\ f_{k,\,l}^{j+3/4} &= 2\varphi_{k,\,l}^{j+1/2}+2\,(\mu_{k,\,l+1}\,v_{k,\,l+1}^{j}-\mu_{k,\,l}\,v_{k,\,l}^{j}). \end{aligned} \tag{6.48}$$

On a donc l'algorithme de calcul suivant pour le problème (6.30).

On définit les champs initiaux des $u_{k,\,l}^{0}$, $v_{k,\,l}^{0}$, $\varphi_{k,\,l}^{0}$, les deux premières fonctions étant assujetties à vérifier l'analogue aux différences de (6.29). On calcule $f_{k,\,l}^{j+1/4}$ par la première formule (6.48) et on résout la première équation aux différences (6.47) avec la condition (6.45) sur la frontière de D_h. On détermine

$$\varphi_{k,\,l}^{j+1/2}=2\varphi_{k,\,l}^{j+1/4}-\varphi_{k,\,l}^{j}$$

par les valeurs obtenues $\varphi_{k,\,l}^{j+1/4}$ à l'aide de la premizère formule (6.46). On résout la seconde équation (6.47) sous la condition (6.45) et on calcule moyennant (6.46) les quantités

$$\varphi_{k,\,l}^{j+1}=2\varphi_{k,\,l}^{j+3/4}-\varphi_{k,\,l}^{j+1/2}.$$

On utilise $\varphi_{k,\,l}^{j+1}$ ainsi obtenus pour déterminer $u_{k,\,l}^{j+1}$ et $v_{k,\,l}^{j+1}$ à l'aide des premières relations (6.42) et (6.43).

La méthode décrite se généralise aisément au cas d'équations hyperboliques plus complexes et elle conduit en général à des schémas absolument stables d'ordre d'approximation 2 sous une condition faible imposée à la définitivité des opérateurs A_α.

CHAPITRE 6

AMÉLIORATION DE LA PRÉCISION DES SOLUTIONS APPROCHÉES AU SENS DE RICHARDSON

La formation des solutions approchées très précises constitue un problème pressant de l'Analyse numérique. Parmi les méthodes connues, les schémas aux différences et les schémas variationnels aux différences de haute précision (Runge-Kutta, prédiction-correction, *extrapolation de Richardson*) trouvent les plus nombreuses applications.

Ce chapitre est essentiellement consacré à une méthode dont l'idée remonte à Richardson qui l'a appelée *extrapolation à la limite* [IV]. Elle consiste à construire les solutions approchées avec une précision donnée moyennant une suite de réseaux et les approximations associées d'un même type. Cette tactique permet de se borner aux approximations standards d'ordre de précision 1 et 2.

L'extrapolation à la limite s'applique aujourd'hui à des problèmes des plus variés, voire non linéaires, dont plusieurs, parmi les plus simples, font l'objet des pages suivantes.

6.1. Equation différentielle ordinaire du premier ordre

Nous allons étudier l'effet de l'extrapolation de Richardson sur un pas d'un réseau de discrétisation dans le cas d'une équation linéaire simple.

Soit $u(t)$ solution de l'équation différentielle

$$u' + a(t)\, u = f(t), \quad t \in (0, 1), \tag{1.1}$$

avec la condition initiale

$$u(0) = u_0. \tag{1.2}$$

On suppose que

$$a(t) \geqslant 0, \quad t \in (0, 1), \tag{1.3}$$

et que toutes les fonctions sont suffisamment régulières.

Divisons le segment [0, 1] par les points « entiers »

$$t_j = j\tau, \quad j = 0, 1, \ldots, M \tag{1.4}$$

en pas de longueur $\tau = 1/M$, M étant un nombre entier, et les points « intermédiaires »

$$t_{j+1/2} = (j + 1/2)\,\tau, \quad j = 0, 1, \ldots, M - 1. \tag{1.5}$$

Conformément à Crank-Nicholson (voir n° 5.2), on remplace l'équation (1.1) relative aux nœuds intermédiaires par le système algébrique approché

$$\frac{u^{j+1} - u^j}{\tau} + a^{j+1/2}\,\frac{u^{j+1} + u^j}{2} = f^{j+1/2}, \quad j = 0, 1, \ldots, M-1. \tag{1.6}$$

On adjoint la condition initiale

$$u^0 = u_0 \tag{1.7}$$

et on définit tous les u^j de façon récurrente :

$$u^j = \left(1 + \frac{\tau}{2} a^{j-1/2}\right)^{-1} \left[\left(1 - \frac{\tau}{2} a^{j-1/2}\right) u^{j-1} - \frac{\tau}{2} f^{j-1/2}\right], \tag{1.8}$$
$$j = 1, 2, \ldots, M.$$

Il est déjà connu que le problème aux différences (1.6)-(1.7) est d'ordre d'approximation 2, et on évalue la stabilité

$$|u^j| \leqslant u_0 + \max_{0 \leqslant j \leqslant M-1} |f^{j+1/2}| \tag{1.9}$$

en raisonnant comme au n° 5.2 et en prenant, au lieu de la matrice Λ^j définie positive, le nombre positif $a^{j+1/2}$. Selon le numéro 1.4, la solution du problème aux différences approche donc la solution du problème différentiel à l'ordre 2 en τ

$$\max_{0 \leqslant j \leqslant M} |u^j - (u)^j| \leqslant C_1 \tau^2, \tag{1.10}$$

la constante C_1 étant indépendante de τ.

Voyons ce qu'il en devient en calcul sur ordinateur dans le cas de l'équation

$$u' + tu = t, \quad t \in (0, 1), \tag{1.11}$$

avec la condition initiale

$$u(0) = 2. \tag{1.12}$$

On vérifie aisément que ce problème admet comme solution exacte

$$u(t) = e^{-\frac{t^2}{2}} + 1. \tag{1.13}$$

L'expérience numérique a consisté à approcher la solution de (1.4)-(1.8) pour $M = 10$, 20, 50, 100, 200 et à définir

$$\xi = \max_{0 \leqslant j \leqslant M} |u^j - (u)^j|. \tag{1.14}$$

La fig. 18 représente ξ en fonction de M ($+\,+$ — valeurs calculées; —— — valeurs théoriques). On voit que les expériences numériques confirment l'estimation (1.10). La vérification ultérieure révèle que (1.10) ne saurait être améliorée en ce qui concerne l'ordre de grandeur, i.e.

$$\xi \geqslant C_2 \tau^2 \tag{1.15}$$

avec une constante positive C_2 indépendante de τ. Ainsi, la plus grande erreur (1.14) sur la solution approchée est de l'ordre de τ^ν, $\nu = 2$.

Richardson a observé, au lieu de l'erreur maximale, l'erreur sur la solution approchée relative à chaque point du réseau et formulé en 1910 l'hypothèse suivante sur son comportement avec τ tendant vers 0:

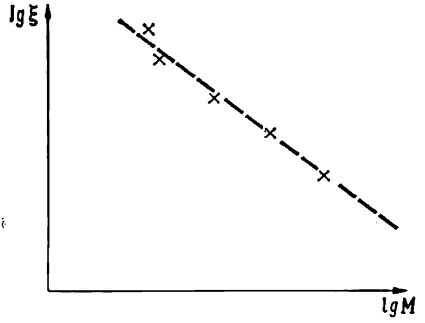


Fig. 18

$$u^j - (u)^j = \tau^2 (v)^j + \eta^j, \quad j = 0, 1, \ldots, M. \tag{1.16}$$

Ceci étant, la partie principale $\tau^2 (v)^j$ de l'erreur est le produit de τ^2 par les valeurs de $v(t)$ indépendante de τ aux nœuds et le reste η^j est $O(\tau^4)$ quel que soit $j = 0, 1, \ldots, M$.

Nous nous proposons de justifier ce résultat pour le problème (1.1)-(1.3).

On procède d'ordinaire en trois étapes. On établit des conditions nécessaires de (1.16), on forme avec un problème aux limites dont $v(t)$ est solution et on démontre enfin la borne de la fonction discrète η^j/τ^4.

Ainsi, si la relation (1.16) a lieu, on écrit

$$u^j = (u)^j + \tau^2 (v)^j + \ldots, \qquad j = 0, 1, \ldots, M, \tag{1.17}$$

où le reste est supposé être $O(\tau^4)$. On fixe j et on porte les valeurs de u^j dans (1.6):

$$\frac{(u)^{j+1} - (u)^j}{\tau} + a^{j+1/2} \frac{(u)^{j+1} + (u)^j}{2} + \\ + \tau^2 \left(\frac{(v)^{j+1} - (v)^j}{\tau} + a^{j+1/2} \frac{(v)^{j+1} + (v)^j}{2} \right) + \ldots = f^{j+1/2}.$$

Développons u et v en série de Taylor

$$(u)^{j+1/2 \pm 1/2} = \left(u \pm \frac{\tau}{2} u' + \frac{\tau^2}{8} u'' \pm \frac{\tau^3}{48} u''' \right)^{j+1/2} + \ldots,$$

$$(v)^{j+1/2 \pm 1/2} = \left(v \pm \frac{\tau}{2} v' \right)^{j+1/2} + \ldots, \tag{1.18}$$

et faisons correspondre à chaque terme une valeur de la fonction au nœud $j + 1/2$:

$$\left(u' + au + \tau^2\left(\frac{1}{24}u''' + \frac{1}{8}au''\right) + \tau^2(v' + av)\right)^{j+1/2} + \ldots = f^{j+1/2}.$$

Les quantités de l'ordre de $O(\tau^0)$ étant éliminées par suite de la validité de (1.1) en $j + 1/2$, il faut pour avoir l'égalité avec $\tau \to 0$ qu'on ait

$$(v' + av)^{j+1/2} = \left(-\frac{1}{24}u''' - \frac{1}{8}au''\right)^{j+1/2}, \quad j = 0, 1, \ldots, M-1. \tag{1.19}$$

Considérons l'égalité (1.17) pour $j = 0$:

$$u^0 = u_0 + \tau(v)^0 + \ldots \tag{1.20}$$

Les conditions initiales (1.2) et (1.7) entraînent $u^0 = u_0$, d'où la condition pour v^0 quand $\tau \to 0$

$$(v)^0 = 0. \tag{1.21}$$

Les conditions nécessaires découlant de (1.17) sont donc construites. On suppose maintenant que $v(t)$ est définie si (1.19) a lieu sur (0, 1) tout entier, i.e. on a

$$v'(t) + a(t)v(t) = -\frac{1}{24}u'''(t) - \frac{1}{8}au''(t), \qquad t \in (0, 1). \tag{1.22}$$

Adjoignons à cette équation (par rapport à $v(t)$) la condition initiale

$$v(0) = 0, \tag{1.23}$$

conséquence de (1.21), il vient un problème différentiel qui possède la solution $v(t)$ suffisamment régulière, v ne dépendant évidemment pas de τ.

On démontre que la fonction discrète η définie par $M + 1$ relations

$$\eta^j = u^j - (u)^j - \tau^2(v)^j \tag{1.24}$$

prend des valeurs en $O(\tau^4)$.

On fixe un j et on porte η dans l'opérateur aux différences du problème (1.6):

$$\frac{\eta^{j+1} - \eta^j}{\tau} + a^{j+1/2}\frac{\eta^{j+1} + \eta^j}{2} = \frac{u^{j+1} - u^j}{\tau} + a^{j+1/2}\frac{u^{j+1} + u^j}{2} +$$
$$+ \left(\frac{(u)^{j+1} - (u)^j}{\tau} + a^{j+1/2}\frac{(u)^{j+1} + (u)^j}{2}\right) -$$
$$- \tau^2\left(\frac{(v)^{j+1} - (v)^j}{\tau} + a^{j+1/2}\frac{(v)^{j+1} + (v)^j}{2}\right). \tag{1.25}$$

On effectue certaines transformations sur le second membre. La formule (1.6) entraîne l'égalité des deux premiers termes et de $f^{j+1/2}$,

et on associe toutes les fonctions des termes restants aux valeurs au nœud $j + 1/2$ par le développement (1.18) avec restes sous forme de Lagrange :

$$\frac{\eta^{j+1}-\eta^j}{\tau}+a^{j+1/2}\,\frac{\eta^{j+1}+\eta^j}{2}=$$

$$=\left(f-u'-au-\tau^2\left(\frac{1}{24}\,u'''+\frac{1}{8}\,au''\right)-\tau^2(v'+av)\right)^{j+1/2}-$$

$$-\tau^4\left(\frac{1}{16\cdot 5!}\,u^{\mathrm{V}}(\xi_j^1)+\frac{1}{16\cdot 4!}\,a^{j+1/2}u^{\mathrm{IV}}(\xi_j^1)\right)-$$

$$-\tau^4\left(\frac{1}{24}\,v'''(\xi_j^3)+\frac{1}{8}\,a^{j+1/2}v''(\xi_j^4)\right),$$

ξ_j^i étant des points de l'intervalle $(t_j,\ t_{j+1})$.

On simplifie la dernière égalité car u est la solution de (1.1) et v celle de (1.22) :

$$\frac{\eta^{j+1}-\eta^j}{\tau}+a^{j+1/2}\,\frac{\eta^{j+1}+\eta^j}{2}=-\tau^4\left(\frac{1}{16\cdot 5!}\,u^{\mathrm{V}}(\xi_j^1)+\right.$$

$$\left.+\frac{1}{16\cdot 4!}\,a^{j+1/2}u^{\mathrm{IV}}(\xi_j^2)+\frac{1}{24}\,v'''(\xi_j^3)+\frac{1}{8}\,a^{j+1/2}v''(\xi_j^4)\right). \qquad (1.26)$$

Si u et v possèdent sur $[0,\ 1]$ des dérivées continues figurant dans (1.26), alors le second membre est une quantité bornée en valeur absolue par $\tau^4 C_3$, la constante C_3 étant indépendante de τ. Utilisons ce résultat valide pour tout $j = 0,\ 1,\ \ldots,\ M$ et l'estimation (1.9), il vient

$$|\,\eta^j\,| \leqslant |\,\eta^0\,| + \tau^4 C_3$$

et la définition de η^j implique $\eta^0 = 0$ par suite de (1.21), (1.7) et (1.23). Donc

$$|\,\eta^j\,| \leqslant \tau^4 C_3, \qquad j = 0,\ 1,\ \ldots,\ M, \qquad (1.27)$$

ce qui prouve l'hypothèse (1.16).

Ce développement est à la base d'une méthode d'amélioration de la précision.

Soient u_τ et $u_{\tau/2}$ solutions de deux problèmes approchés (1.6), (1.7) avec pas respectifs τ et $\tau/2$. Bien que chaque solution soit en $O(\tau^2)$, on génère facilement une solution exacte à l'ordre 4 basée sur un réseau de pas τ. Soit donc $t_j = \tau_i$ un point quelconque d'un réseau de pas τ, auquel cas la combinaison linéaire

$$\bar{u}^j = \frac{4}{3}\,u_\tau^j - \frac{1}{3}\,u_{\tau/2}^{2j} \qquad (1.28)$$

approche la solution exacte u à l'ordre 4 en τ :

$$|\,\bar{u}^j - u(j\tau)\,| \leqslant \tau^4 C_4, \qquad j = 0,\ 1,\ \ldots,\ M, \qquad (1.29)$$

la constante C_4 étant indépendante de τ.

On démontre ce résultat. Les deux solutions approchées se développent au point $j\tau$ en

$$u_\tau^j = u(j\tau) + \tau^2 v(j\tau) + \eta_\tau^j$$

et

$$u_{\tau/2}^{2j} = u(j\tau) + \frac{\tau^2}{4} v(j\tau) + \eta_{\tau/2}^{2j}.$$

Pour former $\bar{u}^j$ on utilise la coïncidence de $u(j\tau)$ et $v(j\tau)$:

$$\bar{u}^j = \left(\frac{4}{3} - \frac{1}{3}\right) u(j\tau) + \tau^2 \left(\frac{4}{3} \cdot \frac{1}{4} - \frac{1}{3}\right) v(j\tau) + \left(\frac{4}{3} \eta_{\tau/2}^{2j} - \frac{1}{3} \eta_\tau^j\right).$$

D'où

$$\bar{u}^j = u(j\tau) + \frac{4}{3} \eta_{\tau/2}^{2j} - \frac{1}{3} \eta_\tau^j.$$

On obtient, compte tenu de la majoration (1.27),

$$\left| \frac{4}{3} \eta_{\tau/2}^{2j} - \frac{1}{3} \eta_\tau^j \right| \leqslant \frac{4}{3} \cdot \frac{\tau^4}{16} C_3 + \frac{\tau^4}{3} C_3 = \frac{5}{12} \tau^4 C_3,$$

d'où (1.29) si $C_4 = \frac{5}{12} C_3$.

Le procédé décrit se transpose aux réseaux non réguliers, mais les raisonnements se compliquent.

Si l'on augmente le nombre de termes du développement (1.16), l'ordre de précision de la solution approchée obtenue est supérieur à 4. Un exemple de cette généralisation est donné au numéro suivant.

6.2. Equation de diffusion à une dimension

Nous allons étudier le problème de Dirichlet pour l'équation de diffusion simple

$$-\frac{d^2u}{dx^2} + q(x)\, u = f(x), \qquad x \in (0,\ 1), \tag{2.1}$$

$$u(0) = 0, \qquad u(1) = 0, \tag{2.2}$$

sous l'hypothèse que

$$q(x) \geqslant 0 \tag{2.3}$$

et que q, f et, partant, u sont suffisamment régulières.

Nous appliquerons au problème (2.1)-(2.2), *primo*, une variante de l'extrapolation à la limite (Marchuk, Shaydourov [IV; 19, 40]) en termes de méthode variationnelle des différences de Galerkin et, *secundo*, l'extrapolation à la limite pour les schémas aux différences qui fournit les solutions approchées en $O(h^{2k})$, $k = 2$, 3 et plus grand, quand $h \to 0$.

6.2.1. Méthode de Galerkin

Contrairement aux méthodes variationnelles des différences de haute précision qui utilisent plus de fonctions de base très régulières (l'exemple en est donné au n° 2.4) que de coutume, la méthode ci-dessous augmente la précision des solutions approchées en prenant des combinaisons linéaires de solutions obtenues avec plusieurs pas différents et en recourant aux fonctions de base linéaires par morceaux.

Pour approcher le problème initial on divise l'intervalle [0, 1] en N parties de longueur $h = 1/N$ par les points équidistants

$$x_k = kh, \qquad k = 0, 1, \ldots, N, \tag{2.4}$$

et on construit les fonctions de base ω_k du point 2.3.2:

$$\omega_k(x) = \begin{cases} 0 & \text{si } x \in (-\infty, x_{k-1}], \\ 1 + \frac{x - x_k}{h} & \text{si } x \in (x_{k-1}, x_k], \\ 1 - \frac{x - x_k}{h} & \text{si } x \in (x_k, x_{k+1}), \\ 0 & \text{si } x \in [x_{k+1}, +\infty), \end{cases}$$
$$k = 1, 2, \ldots, N-1.$$

On multiplie (2.1) par chaque fonction de base et on intègre le résultat, il vient

$$-\int_{x_{k-1}}^{x_{k+1}} \left(\frac{d^2u}{dx^2}\omega_k + qu\omega_k\right) dx = \int_{x_{k-1}}^{x_{k+1}} f\omega_k \, dx,$$
$$k = 1, \ldots, N-1,$$

à condition que $u(x)$ est solution du problème (2.1), (2.2). Intégrons par parties, compte tenu de l'égalité $\omega_k(x_{k\pm1}) = 0$:

$$\int_{x_{k-1}}^{x_{k+1}} \left(\frac{du}{dx}\frac{d\omega_k}{dx} + qu\omega_k\right) dx = \int_{x_{k-1}}^{x_{k+1}} f\omega_k \, dx. \tag{2.5}$$

On introduit les notations

$$(v, w) = \int_0^1 v(x)\, w(x)\, dx, \quad [v, w] = \int_0^1 \left(\frac{dv}{dx}\frac{dw}{dx} + qvw\right) dx$$

destinées à simplifier l'exposé ultérieur. Avec ces notations, les égalités précédentes s'écrivent

$$[u, \omega_k] = (f, \omega_k), \qquad k = 1, 2, \ldots, N-1. \tag{2.6}$$

Ces identités aidant, on cherche une solution approchée $u^h(x)$ de la forme

$$u^h(x) = \sum_{j=1}^{N-1} \alpha_j^h \omega_j(x), \tag{2.7}$$

avec des constantes α_l^h définies par les égalités obtenues à partir de (2.6) par substitution $u^h(x) = u(x)$:

$$\sum_{l=1}^{N-1} \alpha_l^h [\omega_l, \omega_k] = (f, \omega_k), \qquad k = 1, 2, \ldots, N-1. \tag{2.8}$$

On a intérêt à mettre ce système sous forme matricielle

$$A\alpha \equiv \begin{Vmatrix} b_1 & c_1 & 0 & \ldots & 0 & 0 \\ a_2 & b_2 & c_2 & \ldots & 0 & 0 \\ 0 & a_3 & b_3 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & b_{N-2} & c_{N-2} \\ 0 & 0 & 0 & \ldots & a_{N-1} & b_{N-1} \end{Vmatrix} \begin{Vmatrix} \alpha_1^h \\ \alpha_2^h \\ \vdots \\ \alpha_{N-2}^h \\ \alpha_{N-1}^h \end{Vmatrix} = \begin{Vmatrix} g_1 \\ g_2 \\ \vdots \\ g_{N-2} \\ g_{N-1} \end{Vmatrix}, \tag{2.9}$$

d'éléments

$$a_k = -\frac{1}{h} + \int_{x_{k-1}}^{x_k} \left(1 + \frac{x - x_k}{h}\right) \frac{x_k - x}{h} q\, dx = [\omega_{k-1}, \omega_k],$$

$$b_k = \frac{2}{h} + \int_{x_{k-1}}^{x_k} \left(1 + \frac{x - x_k}{h}\right)^2 q\, dx + \int_{x_k}^{x_{k+1}} \left(1 - \frac{x - x_k}{h}\right)^2 q\, dx = [\omega_k, \omega_k],$$

$$c_k = a_{k+1} = [\omega_{k+1}, \omega_k],$$

et

$$g_k = \int_{x_{k-1}}^{x_k} \left(1 + \frac{x - x_k}{h}\right) f\, dx + \int_{x_k}^{x_{k+1}} \left(1 - \frac{x - x_k}{h}\right) f\, dx = (f, \omega_k).$$

Si une fonction n'est pas un polynôme de degré <2, elle est évidemment approchée uniformément sur [0, 1] à l'ordre 2 en h au plus par les fonctions linéaires par morceaux à base de $\omega_i^h(x)$. On énonce donc, au lieu de l'hypothèse de Richardson, la condition : les coefficients α_k^h admettent pour $h \to 0$ le développement

$$\alpha_k^h = u(kh) + h^2 v(kh) + \eta_k^h, \qquad k = 0, 1, \ldots, N, \tag{2.10}$$

où $u(x)$ est solution du problème (2.1), (2.2), $v(x)$ une fonction régulière indépendante de h et η_k^h une fonction discrète bornée uni-

formément en k par une quantité de l'ordre de $O(h^4)$ en valeur absolue :

$$|\eta_k^h| \leqslant C_1 h^4, \tag{2.11}$$

la constante C_1 étant indépendante de h et k.

On procède en trois étapes : 1) on explique le choix de $v(x)$, 2) on cherche cette fonction et 3) on prouve l'estimation (2.11).

S'il y a validité de (2.10), on obtient en conservant les deux termes principaux :

$$\alpha_k^h = u(x_k) + h^2 v(x_k) + O(h^4). \tag{2.12}$$

Portons dans (2.9) :

$$a_k u(x_{k-1}) + b_k u(x_k) + c_k u(x_{k+1}) + \\ + h^2 (a_k v(x_{k-1}) + b_k v(x_k) + c_k v(x_{k+1})) + O(h^4) = g_k, \\ k = 1, 2, \ldots, N-1, \tag{2.13}$$

où $a_1 = c_{N-1} = 0$ pour fixer les idées. Il y a des doutes sur la grandeur de $O(h^4)$ parce que les calculs simples donnent $O(h^3)$. On montrera qu'elle est en réalité inférieure.

Connaissant les valeurs de $u(x)$ et $v(x)$ aux nœuds du réseau, on construit les fonctions continues

$$\tilde{u}(x) = \sum_{l=1}^{N-1} u(x_l)\, \omega_l(x) \quad \text{et} \quad \tilde{v}(x) = \sum_{l=1}^{N-1} v(x_l)\, \omega_l(x),$$

prolongements linéaires par morceaux de $u(x)$ et $v(x)$ respectivement. La forme des coefficients a_k, b_k, c_k, g_k permet d'aboutir non pas à (2.13) mais aux égalités

$$[\tilde{u}, \omega_k] + h^2 [\tilde{v}, \omega_k] + O(h^4) = (f, \omega_k), \qquad k = 1, 2, \ldots, N-1. \tag{2.14}$$

On en déduit facilement en intégrant par parties

$$-\frac{1}{h} u(x_{k-1}) + \frac{2}{h} u(x_k) - \frac{1}{h} u(x_{k+1}) = \\ = \int_0^1 \frac{du}{dx} \frac{d\omega_k}{dx} dx = \int_0^1 \frac{d\tilde{u}}{dx} \frac{d\omega_k}{dx} dx, \tag{2.15}$$
$$k = 1, 2, \ldots, N-1.$$

Les développements tayloriens de $u(x)$ et $v(x)$

$$u(x) = u(x_k) + (x - x_k) \frac{du}{dx}(x_k) + \frac{(x-x_k)^2}{2} \frac{d^2u}{dx^2}(x_k) + O(h^3), \\ q(x) = q(x_k) + (x - x_k) \frac{dq}{dx}(x_k) + \frac{(x-x_k)^2}{2} \frac{d^2q}{dx^2}(x_k) + O(h^3) \tag{2.16}$$

sur chaque $(x_{k-1},\ x_{k+1})$ permettent de prouver l'égalité

$$(\widetilde{qu},\ \omega_k) = (qu,\ \omega_k) + h^2 (w,\ \omega_k) + O(h^3), \tag{2.17}$$

où

$$w(x) = -\frac{1}{12} q(x) \frac{d^2u}{dx^2}(x).$$

En effet, on a

$$\widetilde{u}(x) = \begin{cases} \frac{1}{h} (u(x_k)(x - x_{k-1}) + u(x_{k-1})(x_k - x)), & x \in (x_{k-1},\ x_k), \\ \frac{1}{h} (u(x_k)(x_{k+1} - x) + u(x_{k+1})(x - x_k)), & x \in (x_k, x_{k+1}), \end{cases}$$

si bien que

$$\widetilde{u}(x) = \begin{cases} u(x_k) + (x_k - x) u'(x_k) + (x_k - x) \frac{h}{2} u''(x_k) + O(h^3) \\ \qquad\qquad \text{pour} \quad x \in (x_{k-1},\ x_k), \\ u(x_k) + (x - x_k) u'(x_k) + (x - x_k) \frac{h}{2} u''(x_k) + O(h^3) \\ \qquad\qquad \text{pour} \quad x \in (x_k,\ x_{k+1}). \end{cases}$$

Utilisons la dernière égalité et les développements (2.16) et intégrons les polynômes sous $\int$, il vient

$$(\widetilde{qu},\ \omega_k) = hu(x_k) q(x_k) -$$
$$- h^3 \left(\frac{1}{12} u(x_k) q''(x_k) + \frac{1}{12} u'(x_k) q(x_k) + \frac{1}{12} u''(x_k) q(x_k) \right) + O(h^3),$$

$$(qu,\ \omega_k) = hu(x_k) q(x_k) - h^3 \frac{1}{12} (u(x_k) q(x_k))'' + O(h^4),$$

$$h^2 (w,\ \omega_k) = h^3 w(x_k) + O(h^4).$$

On prouve (2.17) par identification des coefficients des termes en h et h^3.

On réunit (2.15) et (2.17) :

$$[\widetilde{u},\ \omega_k] = [u,\ \omega_k] + h^2 (w,\ \omega_k) + O(h^4). \tag{2.18}$$

S'agissant de $v(x)$ et de (2.16) tronqué, on a par des calculs analogues

$$h^2 (\widetilde{v},\ \omega_k) = h^2 (v,\ \omega_k) + O(h^4).$$

Cette égalité devient par application de (2.15) juste pour $v(x)$

$$h^2 [\widetilde{v},\ \omega_k) = h^2 [v,\ \omega_k] + O(h^4). \tag{2.19}$$

On effectue enfin des substitutions dans le premier membre de (2.14) à l'aide de (2.18) et (2.19):

$$[u,\ \omega_k] + h^2\,(w,\ \omega_k) + h^2\,[v,\ \omega_k] + O\,(h^4) = (f,\ \omega_k), \qquad (2.20)$$
$$k = 1,\ 2,\ \ldots,\ N-1.$$

On veut que l'identité a lieu pour tout h et on exige donc la coïncidence des coefficients principaux.

On identifie les coefficients des termes en h^0 et h^2:

$$[u,\ \omega_k] = (f,\ \omega_k), \qquad (2.21)$$
$$[v,\ \omega_k] = -(w,\ \omega_k), \qquad (2.22)$$
$$k = 1,\ 2,\ \ldots,\ N-1.$$

Les conditions nécessaires sont donc établies. Le premier groupe de ces égalités coïncide avec (2.6), conséquence de (2.1), et il suffit, pour avoir les égalités suivantes, de prendre $v\,(x)$ pour solution du problème aux limites

$$-\frac{d^2v}{dx^2} + qv = -w\,(x), \qquad (2.23)$$

$$v\,(0) = v\,(1) = 0. \qquad (2.24)$$

Ainsi, on a établi l'équation qui définit $v\,(x)$. Il reste à démontrer la majoration (2.11) de la fonction discrétisée η_k^h définie par

$$\eta_k^h = \alpha_k^h - u\,(x_k) - h^2 v\,(x_k), \qquad k = 0,\ 1,\ \ldots,\ N. \qquad (2.25)$$

On reconstitue, à partir des valeurs de η_l^h aux nœuds du réseau, le prolongement linéaire par morceaux

$$\widetilde{\eta}^h\,(x) = \sum_{l=1}^{N-1} \eta_l^h \omega_l\,(x),$$

auquel cas (2.25) se récrit

$$\widetilde{\eta}^h\,(x) = u^h\,(x) - u\,(x) - h^2 v\,(x),$$

ce qui entraîne

$$[\widetilde{\eta}\,(x),\ \omega_k] = [u^h,\ \omega_k] - [\widetilde{u},\ \omega_k] - h^2\,[\widetilde{v},\ \omega_k]. \qquad (2.26)$$

Avec u et q supposées avoir sur $[0,\ 1]$ des dérivées continues d'ordre 1, 2, 3 et 4, on a le développement taylorien de u:

$$u\,(x) = \sum_{l=0}^{3} \frac{d^l u}{dx^l}\,(x_k)\,\frac{(x-x_k)^l}{l!} + \frac{(x-x_k)^4}{4!}\,\frac{d^4 u}{dx^4}\,(\xi_k^x), \qquad x \in (x_{k-1},\ x_{k+1}),$$

un développement analogue de q et

$$v\,(x) = v\,(x_k) + (x-x_k)\,\frac{dv}{dx}\,(x_k) + \frac{(x-x_k)^2}{2}\,\frac{d^2 v}{dx^2}\,(\rho_k^x), \qquad x \in (x_{k-1},\ x_{k+1}),$$

avec ξ_k^x et $\rho_k^x \in (x_{k-1}, x_{k+1})$, les deux derniers développements comptant deux termes de moins.

Par les mêmes raisonnements que dans le cas de conditions nécessaires, on ramène (2.26) à

$$[\tilde{\eta}, \omega_k] = \sigma_k, \quad |\sigma_k| \leqslant C_2 h^5, \tag{2.27}$$

du moment que le choix de $v(x)$ pour solution du problème (2.23) a fait simplifier tous les termes en h et h^3.

On multiplie toutes les égalités (2.27) par η_l^h et on somme par rapport à tous les l :

$$[\tilde{\eta}^h, \tilde{\eta}^h] = \sum_{l=1}^{N-1} \sigma_l \eta_l^h.$$

On substitue à σ_l et à η_l^h les valeurs maximales

$$[\tilde{\eta}^h, \tilde{\eta}^h] \leqslant N C_2 h^5 \max_{1 \leqslant j \leqslant N-1} |\eta_j^h| = C_2 h^4 \max_{0 \leqslant x \leqslant 1} |\tilde{\eta}^h(x)|. \tag{2.28}$$

Pour obtenir la minoration on fait appel à l'inégalité suivante qui découle de l'inégalité de Cauchy-Bouniakovski pour $y(x) = 0$ en $x = 0$:

$$|y(x)| = \left| \int_0^x y'(t)\, dt \right| \leqslant \left(\int_0^x 1\, dt \right)^{1/2} \left(\int_0^x (y'(t))^2\, dt \right)^{1/2} \leqslant$$

$$\leqslant \sqrt{x} \left(\int_0^1 (y'(t))^2\, dt \right)^{1/2} \leqslant \left(\int_0^1 \{(y'(t))^2 + q(y(t))^2\}\, dt \right)^{1/2}.$$

Comme $\tilde{\eta}^h(x)$ est nulle au point $x = 0$, on a

$$|\tilde{\eta}^h(x)|^2 \leqslant [\tilde{\eta}^h, \tilde{\eta}^h], \qquad x \in [0, 1],$$

et l'estimation (2.28) implique

$$\max_{0 \leqslant x \leqslant 1} |\tilde{\eta}^h(x)|^2 \leqslant [\tilde{\eta}^h, \tilde{\eta}^h] \leqslant C_2 h^4 \max_{0 \leqslant x \leqslant 1} |\tilde{\eta}^h(x)|,$$

et finalement

$$\max_{0 \leqslant x \leqslant 1} |\tilde{\eta}^h(x)| = \max_{1 \leqslant l \leqslant N} |\tilde{\eta}_l^h(x)| \leqslant C_2 h^4, \tag{2.29}$$

ce qui justifie (2.10).

Les raisonnements suivants imitent ceux du numéro précédent, à savoir on forme deux réseaux réguliers de pas respectifs h et $h/2$ et on cherche les solutions associées α^h et $\alpha^{h/2}$ des problèmes (2.9). Etant donné un point $x_k = kh$, la solution est fournie au quatrième ordre près par la combinaison linéaire

$$\bar{u}_k = \frac{4}{3} \alpha_{2k}^{h/2} - \frac{1}{3} \alpha_k^h. \tag{2.30}$$

On reprend la démonstration faite pour les équations différentielles ordinaires du premier ordre et on aboutit à l'estimation

$$|\bar{u}_k - u(x_k)| \leqslant \frac{5}{12} C_2 h^4. \tag{2.31}$$

La méthode décrite se généralise au cas de coefficients variables, d'équations quasi linéaires et à d'autres problèmes aux limites.

6.2.2. Méthode aux différences

Soit le problème (2.1), (2.2). Voyons divers aspects du développement en série suivant les puissances de h^2 qui permet d'augmenter l'exactitude des termes en h jusqu'à l'ordre 6 et plus.

Construisons le réseau régulier

$$x_k = kh, \qquad k = 0, 1, \ldots, N,$$

de pas $h = 1/N$ et associons à (2.1), (2.2) le problème approché d'ordre 2

$$\frac{-u_{k-1} + 2u_k - u_{k+1}}{h^2} + q_k u_k = f_k, \qquad k = 1, 2, \ldots, N-1,$$

$$u_0 = u_N = 0. \tag{2.32}$$

On fait l'hypothèse suivante sur sa solution pour $h \to 0$:

$$u_k = \sum_{j=0}^{l-1} h^{2j} v_j(x_k) + \eta_k, \qquad k = 0, 1, \ldots, N. \tag{2.33}$$

Ici $v_j(x)$ sont des fonctions régulières indépendantes de h et η_k^h est une fonction discrète dont le module est borné uniformément en h par une quantité en h^{2l}:

$$|\eta_k| \leqslant C_1 h^{2l}, \tag{2.34}$$

C_1 étant une constante indépendante de h et k.

Dans un premier temps, on utilise la formule de Taylor

$$v_j(x_{k\pm1}) = \sum_{r=0}^{2l+1-2j} \frac{(\pm h)^r}{r!} \frac{d^r v_j}{dx^r}(x_k) + O(h^{2l+2-2j}). \tag{2.35}$$

Portons le développement (2.33) dans le premier membre du système (2.32), il vient, compte tenu de (2.35),

$$\sum_{j=0}^{l-1} h^{2j} \left\{ \sum_{r=0}^{l-j} -\frac{2h^{2r}}{(2r+2)!} \frac{d^{2r+2} v_j}{dx^{2r+2}}(x_k) + q(x_k) v_j(x_k) \right\} + O(h^{2l}) = f_k.$$

Réduisons les termes semblables:

$$\sum_{j=0}^{l-1} h^{2j}\Big(\sum_{r=0}^{j} -\frac{2}{(2j-2r+2)!}\,\frac{d^{2j-2r+2}v_r(x_k)}{dx^{2j-2r+2}} + \\ + q(x_k)\,v_j(x_k)\Big) + O(h^{2l}) = f_k, \quad (2.36)$$

$$k = 1, 2, \ldots, N-1.$$

La substitution immédiate des développements n'entraîne évidemment pas la propriété, pour le reste de (2.36), d'être de la forme $O(h^{2l})$. Cette propriété sera justifiée plus loin. Egalons les coefficients des termes en h^0, h^2, et ainsi de suite, il vient le système d'égalités

$$-\frac{d^2v_0}{dx^2}(x_k) + q(x_k)\,v_0(x_k) = f_k;$$

$$-\frac{d^2v_j}{dx^2}(x_k) + q(x_k)\,v_j(x_k) - \sum_{r=0}^{j-1} \frac{2}{(2j-2r+2)!}\,\frac{d^{2j-2r+2}v_r}{dx^{2j-2r+2}}(x_k) = 0,$$

$$j = 1, 2, 3, \ldots, l-1, \qquad k = 1, 2, \ldots, N-1.$$

Etant donnée la condition aux limites homogène (2.2), le développement (2.33) donne pour les points $x_0 = 0$ et $x_N = 1$:

$$v_j(x_0) = v_j(x_N) = 0, \qquad j = 0, 1, \ldots, l-1. \quad (2.37)$$

Dans un deuxième temps, on cherche les fonctions $v_j(x)$ successives en tant que solutions du problème de Dirichlet pour une même équation mais avec seconds membres différents:

$$\frac{d^2v_0}{dx^2} + qv_0 = f, \qquad x \in (0, 1),$$
$$v_0(0) = v_0(1) = 0; \quad (2.38)$$

$$-\frac{d^2v_j}{dx^2} + qv_j = \sum_{r=0}^{j-1} \frac{2}{(2j-2r+2)!}\,\frac{d^{2j-2r+2}v_r}{dx^{2j-2r+2}}, \qquad x \in (0, 1),$$
$$v_j(0) = v_j(1) = 0, \quad (2.39)$$
$$j = 1, 2, \ldots, l-1.$$

On vérifie aisément que $v_0(x) \equiv u(x)$, solution du problème (2.1), (2.2).

On définit maintenant la fonction discrète η^h par les égalités

$$\eta_k^h = u_k^h - \sum_{j=0}^{l-1} h^{2j} v_j(x_k), \qquad k = 0, 1, \ldots, N,$$

et on passe à la démonstration de (2.34). On porte η^h dans l'opérateur aux différences du problème (2.32) et on utilise la formule de Taylor (2.35) avec reste sous forme de Lagrange

$$\frac{-\eta^h_{k-1}+2\eta^h_k-\eta^h_{k+1}}{h^2}+q_k\eta^h_k=f_k+$$
$$+\sum_{j=0}^{l-1}h^{2j}\left\{\sum_{r=0}^{j}\frac{2}{(2j-2r+2)!}\frac{d^{2j-2r+2}v_r}{dx^{2j-2r+2}}(x_k)-\right.$$
$$\left.-q(x_k)\,v_j(x_k)\right\}+h^{2l}\varepsilon^h_k. \qquad (2.40)$$

On a

$$\varepsilon^h_k=\sum_{r=0}^{j}\frac{2}{(2l-2r+2)!}\frac{d^{2l-2r+4}v_r}{dx^{2l-2r+4}}(\xi_{j,k})$$

avec $\xi_{j,k}\in(0,1)$, si bien qu'une condition suffisante de borne uniforme $|\varepsilon^h_k|\leqslant C_2$ est que $q(x)$ et $f(x)$ possèdent sur $[0,1]$ $2l$ dérivées continues, propriété qui garantit la régularité requise de toutes les fonctions $v_j(x)$ admettant sur $[0,1]$ $2l-2j+2$ dérivées continues chacune.

Etant donné le procédé d'obtention des $v_j(x)$, on simplifie plusieurs termes du second membre de (2.40):

$$\frac{-\eta^h_{k-1}+2\eta^h_k-\eta^h_{k+1}}{h^2}+q_k\eta^h_k=h^{2l}\varepsilon^h_k, \qquad k=1,2,\ldots,N-1. \qquad (2.41)$$

Les fonctions η^h vérifient de plus les égalités

$$\eta^h_0=\eta^h_N=0, \qquad (2.42)$$

conséquence de la façon dont η^h est définie aux nœuds frontières et des conditions aux limites des problèmes (2.32), (2.38) et (2.39).

Selon le numéro 2.1, il y a possibilité du système (2.41), (2.42). On démontre l'estimation de la solution

$$\max_{0\leqslant k\leqslant N}|\widetilde{\eta}^h_k|\leqslant C_2h^{2l}. \qquad (2.43)$$

En multipliant chaque équation (2.41) par η^h_k et en additionnant, on a l'égalité

$$\frac{1}{h^2}\sum_{k=1}^{N}\left\{(\eta^h_k-\eta^h_{k-1})^2+q_k(\eta^h_k)^2\right\}=h^{2l}\sum_{k=1}^{N-1}\varepsilon^h_k\eta^h_k.$$

On supprime les termes positifs $q_k(\eta^h_k)^2$ et on substitue à ε^h_k,η^h_k des valeurs supérieures, ce qui donne

$$\frac{1}{h^2}\sum_{k=1}^{N}(\eta^h_k-\eta^h_{k-1})^2\leqslant C_2h^{2l-1}\max_{0\leqslant k\leqslant N}|\eta^h_k|. \qquad (2.44)$$

On obtient la minoration à l'aide des inégalités (mises sous forme de différences) du point précédent qui sont justes pour toute fonction discrète y nulle au nœud $x_0 = 0$

$$| y_k | = \Big| \sum_{j=1}^{k} (y_j - y_{j-1}) \Big| \leqslant k^{1/2} \Big(\sum_{j=1}^{k} (y_j - y_{j-1})^2 \Big)^{1/2} \leqslant$$

$$\leqslant N^{1/2} \Big(\sum_{j=1}^{N} (y_j - y_{j-1})^2 \Big)^{1/2}.$$

Etant donné $\eta_0^h = 0$, on en tire en posant $y_k = \eta_k^h$:

$$\frac{1}{h} \max_{0 \leqslant k \leqslant N} | \eta_k^h |^2 \leqslant \frac{1}{h^2} \sum_{k=1}^{N} (\eta_k^h - \eta_{k-1}^h)^2.$$

Cette inégalité jointe à (2.44) donne (2.43) au facteur $\frac{1}{h} \max_{0 \leqslant k \leqslant N} \times$ $\times | \eta_k^h |$ près.

Ainsi, on vient d'établir (2.34) avec $C_1 = C_2$ et le développement (2.33) se trouve justifié.

Voyons maintenant comment obtenir des solutions d'ordre de précision h^{2l}. Soit x un point d'intersection de l réseaux de pas respectifs $h_1, h_2, \ldots, h_l$. On associe à chaque réseau un problème approché (2.32) et on cherche la solution u^{hj}. On définit donc en x l valeurs approchées $u^{hj}(x)$. Si les nombres γ_j sont solution du système

$$\sum_{j=1}^{l} \gamma_j = 1, \quad \sum_{j=1}^{l} h_j^{2s} \gamma_j = 0, \qquad s = 1, 2, \ldots, l-1, \qquad (2.45)$$

la solution est donnée à $2l$ près par la formule

$$\bar{u}(x) = \sum_{j=1}^{l} \gamma_j u^{hj}(x). \qquad (2.46)$$

En effet, le report de l développements dans (2.46) donne au lieu de u^{hj}

$$\bar{u}(x) = v_0(x) \sum_{j=1}^{l} \gamma_j + \sum_{s=1}^{l-1} \Big(v_s(x) \sum_{j=1}^{l} h_j^{2s} \gamma_j \Big) + \sum_{j=1}^{l} \gamma_j \eta^{hj}(x),$$

égalité qui se simplifie:

$$\bar{u}(x) = v_0(x) + \sum_{j=1}^{l} \gamma_j \eta^{hj}(x)$$

parce que γ_j sont solution du système (2.45). Comme $v_0 \equiv u$, on a, compte tenu de l'estimation de $\eta^h{}_j(x)$:

$$|\bar{u}(x) - u(x)| \leqslant \sum_{j=1}^{l} |\gamma_j| C_1 h_j^{2l}. \tag{2.47}$$

Cette inégalité ne suffit cependant pas pour démontrer le résultat de précision parce que $|\gamma_j|$ peuvent s'avérer comparables avec les puissances positives de N. Un autre problème obscur est l'existence pour (2.45). On trouve la réponse aux deux questions si l'on recourt aux déterminants de Vandermonde

$$v(\mu_1, \mu_2, \ldots, \mu_l) = \begin{Vmatrix} 1 & 1 & \ldots & 1 \\ \mu_1 & \mu_2 & \ldots & \mu_l \\ \ldots & \ldots & \ldots & \ldots \\ \mu_1^{l-1} & \mu_2^{l-1} & \ldots & \mu_l^{l-1} \end{Vmatrix}.$$

Il est connu que

$$v(\mu_1, \mu_2, \ldots, \mu_l) = \prod_{1 \leqslant i < j \leqslant l} (\mu_j - \mu_i). \tag{2.48}$$

La matrice du système (2.45) a pour déterminant le déterminant de Vandermonde $v(h_1^2, h_2^2, \ldots, h_l^2)$ qui est différent de 0 si tous les h_i sont deux à deux distincts (voir (2.48)). On applique la règle de Cramer au système (2.45):

$$\gamma_j = \frac{v(h_1^2, \ldots, h_{j-1}^2, 0, h_{j+1}^2, \ldots, h_l^2)}{v(h_1^2, \ldots, h_{j-1}^2, h_j^2, h_{j+1}^2, \ldots, h_l^2)},$$

d'où

$$|\gamma_j| = \prod_{1 \leqslant i < j} \frac{h_i^2}{h_j^2 - h_i^2} \prod_{j < i \leqslant l} \frac{h_i^2}{h_i^2 - h_j^2}. \tag{2.49}$$

On suppose que h_j forment une suite croissante et que

$$\frac{h_{i+1}}{h_i} \geqslant 1 + C_3 \tag{2.50}$$

avec la constante $C_3 > 0$. Evidemment,

$$\frac{h_j^2}{h_i^2} \geqslant 1 + 2C_3 \qquad \text{pour } j > i.$$

Donc

$$\frac{h_i^2}{h_j^2 - h_i^2} = \frac{1}{\frac{h_j^2}{h_i^2} - 1} \leqslant \frac{1}{2C_3}, \qquad \text{si } i < j$$

et

$$\frac{h_i^2}{h_i^2-h_j^2}=\frac{1}{1-\frac{h_j^2}{h_i^2}}\leqslant\frac{1}{1-\frac{1}{1+2C_3}}=1+\frac{1}{2C_3},\quad\text{dans le cas contraire.}$$

On simplifie (2.49) moyennant ces inégalités:

$$|\gamma_j|\leqslant\left(1+\frac{1}{2C_3}\right)^{l-1},\tag{2.51}$$

ce qui démontre la borne de tous les $|\gamma_j|$.

Si l'on est donc dans la condition (2.50), il y a existence et unicité de l'ensemble nécessaire des γ_j et l'erreur sur la solution approchée (2.46) s'évalue par

$$|\bar{u}(x)-u(x)|\leqslant\sum_{j=1}^{l}C_1\left(1+\frac{1}{2C_3}\right)^{l-1}h_j^{2l}.\tag{2.52}$$

La démonstration s'achève sur cette inégalité qui continue (2.47) compte tenu de (2.51).

Le cas où les réseaux impliqués sont plus de 2 présente un inconvénient sérieux. Le fait est qu'on n'a cherché jusqu'à présent les solutions très précises qu'aux nœuds communs à tous les réseaux de discrétisation, qui peuvent être peu nombreux. En effet, on n'en compte que N même dans la situation la plus favorable (réseaux à pas h, $h/2$, . . ., h/l). L'interpolation à l'aide des fonctions splines, des polynômes de Newton et d'Hermite évite la perte de nombreuses valeurs calculées et fournit des solutions approchées très précises. Plaçons-nous dans un cas fort simple et utilisons les polynômes d'interpolation de Newton.

Soit t un point quelconque de l'intervalle fermé [0, 1]. On fixe n'importe quel j et on considère un réseau de pas h_j. Les nœuds les plus proches de t sont en nombre $2l$. Connaissant les valeurs de u^{hj} en ces nœuds, on construit le polynôme d'interpolation de Newton de degré $2l-1$. Soit $P_j(x;\ u^{hj})$ ce polynôme. Le développement (2.33) entraîne

$$P_j(x;\ u^{hj})=\sum_{k=0}^{l-1}h^{2k}P_k(x;\ v_k)+P_j(x;\ \eta^h).$$

Le polynôme interpole les valeurs des fonctions régulières à $O(h^{2l})$ près et ses coefficients sont indépendants de h (Berezin, Zhidkov [III]), si bien que

$$P_j(t;\ u^{hj})=\sum_{k=0}^{l-1}h^{2k}v_k(t)+O(h_j^{2l}).\tag{2.53}$$

Le développement (2.53) étant valable pour $j = 1, \ldots, l$ permet d'affiner les solutions

$$\bar{u}(t) = \sum_{j=1}^{l} \gamma_j P_j(t;\ u^{hj}) \tag{2.54}$$

avec les mêmes poids γ_j de (2.45). Dans ce cas

$$|\bar{u}(t) - u(t)| \leqslant \sum_{j=1}^{l} h_j^{2l} |\gamma_j| C_4,$$

la constante C_4 comprenant les constantes de la formule d'interpolation et les erreurs sur la solution relative aux nœuds. La démonstration s'effectue comme dans le cas de réseaux ayant un point en commun.

6.3. Méthode de décomposition pour un problème d'évolution

Nous allons étudier le problème différentiel

$$\frac{du}{dt} + A(t)\, u = f(t), \qquad t \in (0,\ 1),$$
$$u(0) = u_0, \tag{3.1}$$

où $A(t)$ est une matrice (n, n) formée de fonctions et $u(t)$ et $f(t)$ sont des fonctions vectorielles de n composantes. Supposons que A admet la représentation

$$A(t) = A_1(t) + A_2(t),$$

avec $A_i(t)$ des matrices semi-définies positives pour tout $t \in [0, 1]$.

Partageons le segment $[0, 1]$ par $M + 1$ points équidistants

$$t_j = j\tau, \qquad j = 0, 1, \ldots, M, \tag{3.2}$$

la longueur du pas étant $\tau = 1/M$. Considérons le schéma de décomposition implicite

$$\begin{gathered}(E + \tau A_1(t_j))\, u^{j-1/2} = u^{j-1} + \tau f^j, \\ (E + \tau A_2(t_j))\, u^j = u^{j-1/2}, \\ j = 1,\ 2,\ \ldots,\ M, \qquad u^0 = u_0,\end{gathered} \tag{3.3}$$

avec E la matrice unité.

On munit les vecteurs de dimension n de la norme

$$\|v\| = \Big(\sum_{i=1}^{n} v_i^2\Big)^{1/2}.$$

Результаты пункта 4.4.3 влекут оценку

$$\max_{0\leqslant j\leqslant M}\|u^j\|\leqslant\|u_0\|+\max_{0\leqslant j\leqslant M}\|f^j\| \tag{3.4}$$

характеризующую устойчивость схемы.

Мы намерены обосновать разложение

$$u^j=\sum_{k=0}^{l-1}\tau^k v_k(t_j)+\eta_\tau^j,\qquad j=0,\ 1,\ \ldots,\ M \tag{3.5}$$

с векторными функциями v_k, регулярными и независимыми от τ, и векторной дискретной функцией η_τ, удовлетворяющей неравенству

$$\|\eta_\tau^j\|\leqslant C_1\tau^l, \tag{3.6}$$

где постоянная C_1 не зависит от τ и k.

Система (3.3) после исключения промежуточных значений $u^{j-1/2}$ записывается в виде

$$(E+\tau A_1(t_j))(E+\tau A_2(t_j))u^j=u^{j-1}+\tau f^j. \tag{3.7}$$

Предположим, что разложение (3.5) существует, и подставим его в (3.7) вместо u^j, получим

$$(E+\tau A_1(t_j))(E+\tau A_2(t_j))\sum_{k=0}^{l-1}\tau^k v_k(t_j)=$$

$$=\sum_{k=0}^{l-1}\tau^k v_k(t_{j-1})+O(\tau^{l+1})+\tau f^j. \tag{3.8}$$

Разложим каждое v_k в ряд Тейлора в окрестности точки t_j:

$$(v_k)^{j-1}=\sum_{i=0}^{j-k-1}\frac{d^i(v_k)^j}{dt^i}\frac{(-\tau)^i}{i!}+\ \ldots,$$

приведем подобные члены в правой части (3.8) и произведем умножение сумм в левой части:

$$(v_0+\tau(v_1+Av_0))+$$

$$+\sum_{j=2}^{l-1}\tau^j(v_j+Av_{j-1}+A_1A_2v_{j-2})+\tau^l(Av_{l-1}+A_1A_2v_{l-2})^j=$$

$$=\sum_{k=0}^{l-1}\tau^k\sum_{i=0}^{k}\frac{(-1)^{k-i}}{(k-i)!}\frac{d^{k-i}(v_i)^j}{dt^{k-i}}+\tau f^i+O(\tau^{l+1})+\tau^l\sum_{i=0}^{l-1}\frac{(-1)^{l-i}}{(l-i)!}\frac{d^{l-i}(v_i)^j}{dt^{l-i}}.$$

Cette égalité est juste pour tout τ, si bien qu'on égale les coefficients dominants:

$$(v_0)^j = (v_0)^j,$$

$$(v_1 + Av_0)^j = \left(-\frac{dv_0}{dt} + v_1\right)^j + f^j;$$

$$(v_k + Av_{k-1} + A_1A_2v_{k-2})^j = \sum_{i=0}^{k} \frac{(-1)^{k-i}}{(k-i)!} \frac{d^{k-i}(v_i)^j}{dt^{k-i}},$$

$$k = 2,\ 3,\ \ldots,\ l-1;$$

$$(Av_{l-1} + A_1A_2v_{l-2})^j = \sum_{i=0}^{l-1} \frac{(-1)^{l-i}}{(l-i)!} \frac{d^{l-i}(v_k)^j}{dt^{l-i}}.$$

On rejette la première égalité triviale, on simplifie par v_k et on change l'ordre des sommations:

$$\left(\frac{dv_0}{dt} + Av_0\right)^j = f^j;$$

$$\left(\frac{dv_k}{dt} + Av_k\right)^j = (-A_1A_2v_{k-1})^j + \sum_{i=0}^{k-1} \frac{(-1)^{k-i}}{(k-i)!} \frac{d^{k-i}(v_i)^j}{dt^{k-i}}, \tag{3.9}$$

$$k = 1,\ 2,\ \ldots,\ l-1.$$

On déduit les égalités qui découlent de la condition aux limites $u = u_0$ et du développement (3.5) au point t_0:

$$u_0 = u = v_0(t_0),\quad v_k(t_0) = 0,\qquad k = 1,\ \ldots,\ l-1. \tag{3.10}$$

Généralisons (3.9) à tout l'intervalle (0, 1) et adjoignons les conditions aux limites (3.10). On a

$$\frac{dv_0}{dt} + Av_0 = f,\qquad t \in (0,\ 1),$$
$$v_0(0) = u_0, \tag{3.11}$$

et

$$\frac{dv_k}{dt} + Av_k = -A_1A_2v_{k-1} + \sum_{i=0}^{k-1} \frac{(-1)^{k-i}}{(k-i)!} \frac{d^{k-i}v_i}{dt^{k-i}},\qquad t \in (0,\ 1),$$
$$v_k(0) = 0, \tag{3.12}$$

$$k = 1,\ 2,\ \ldots,\ l-1.$$

On définit les fonctions v_k moyennant le système obtenu, à savoir on suppose que $v_k(t)$ sont les solutions des problèmes d'évolution qu'on a obtenues de proche en proche à partir de (3.11), (3.12).

La fonction $v_0(x)$ coïncide évidemment avec la solution $u(x)$ du problème (3.1), (3.2).

On définit la fonction vectorielle discrète η_τ par l'égalité

$$\eta_\tau^j = u^j - \sum_{k=0}^{l-1} \tau^k (v_k)^j, \qquad j = 0, 1, \ldots, M,$$

et on prouve (3.6). A cet effet, on substitue à u^j de (3.7) la somme $\sum \tau^k (v_k)^j + \eta_\tau^j$ et on développe les composantes des vecteurs v_k en série de Taylor avec reste sous forme de Lagrange

$$(E+\tau A_1)(E+\tau A_2)\left(\eta_\tau^j + \sum_{k=0}^{l-1} \tau^k (v_k)^j\right) =$$
$$= \sum_{k=0}^{l} \tau^k \sum_{i=0}^{k-1} \frac{(-1)^{k-i}}{(k-i)!} \frac{d^{k-i}(v_i)^j}{dt^{k-i}} + \sum_{k=0}^{l-1} \tau^k (v_k)^j + \tau f^j + \eta_\tau^{j-1} + \varepsilon_\tau^j. \tag{3.13}$$

Si ces vecteurs ont leurs composantes suffisamment régulières, le reste est borné en norme

$$\| \varepsilon_\tau^j \| \leqslant C_2 \tau^{l+1} \tag{3.14}$$

quels que soient τ et j.

Si l'on prend en considération le procédé de génération des v_k, on passe de (3.13) à

$$(E+\tau A_1(t_j))(E+\tau A_2(t_j))\,\eta_\tau^j = \eta_\tau^{j-1} + \varepsilon_\tau^j$$

qui entraîne de suite en raison de l'estimation (3.4) caractéristique de la stabilité

$$\max_{0\leqslant j\leqslant M} \| \eta_\tau^j \| \leqslant \max_{0\leqslant j\leqslant M} \left\| \frac{\varepsilon_\tau^j}{\tau} \right\| \leqslant C_2 \tau^l$$

et on a (3.6), où $C_1 = C_2$.

Le développement (3.5) ainsi démontré permet, comme au numéro 6.2, d'augmenter la précision à $O(\tau^l)$ près à la différence qu'on détermine les poids γ_j de la combinaison linéaire à l'aide du système

$$\sum_{j=1}^{l} \gamma_j = 1, \qquad \sum_{k=1}^{l} \tau_k^s \gamma_k = 0, \qquad s = 1, 2, \ldots, l-1. \tag{3.15}$$

Ce résultat se transpose évidemment au cas $A(t) = \sum_{i=1}^{n} A_i(t)$, si bien qu'on indique un autre schéma aux différences (3.3) très efficace en ce qui concerne le problème quasi linéaire

$$\frac{du}{dt} + A(t, u)\,u = f(t, u), \qquad t \in (0, 1), \quad u(0) = u_0. \tag{3.16}$$

Cette modification s'écrit

$$
\begin{gathered}
(E+\tau A_1(t_j,\ u^j))\,u^{j+1/n}=u^j+\tau f(t_j,\ u^j),\\
(E+\tau A_2(t_j,\ u^j))\,u^{j+2/n}=u^{j+1/n},\\
\dots\dots\dots\dots\dots\dots\\
(E+\tau A_n(t_j,\ u^j))\,u^{j+1}=u^{j+(n-1)/n}\\
j=0,\ 1,\ \dots,\ M-1,\\
u^0=u_0.
\end{gathered}
\tag{3.17}
$$

On démontre le développement (3.5) comme plus haut et on utilise la propriété de semi-définie positivité des matrices $A_i(t, u)$ et le fait que les coefficients et la solution du problème (3.16) présentent une régularité suffisante. La solution approchée de haute précision est également cherchée sous forme de combinaison linéaire (avec γ_j obtenus à partir de (3.15)) de solutions de plusieurs problèmes (3.17) pour les pas τ_k multiples différents.

6.4. Extrapolation de Richardson pour des problèmes à plusieurs variables

S'agissant des cas multidimensionnels, l'extrapolation avec les schémas aux différences simples se heurte à des difficultés. Voyons ces difficultés de plus près et montrons comment on peut les supprimer.

Prenons comme modèle l'équation de Poisson de dimension deux

$$\Delta u = f \quad \text{dans} \quad D \tag{4.1}$$

avec la condition aux limites

$$u = 0 \quad \text{sur} \quad \partial D. \tag{4.2}$$

Si la frontière ∂D est curviligne non régulière, on n'arrive pas, avec les analogues discrétisés simples, à développer l'erreur d'approximation au voisinage de ∂D suivant le pas du réseau, contrairement à ce qu'on a pour l'intérieur du domaine. En effet, la représentation

$$u^h(x) = u(x) + h^2 v_1(x) + h^4 v_2(x) + \dots \tag{4.3}$$

est naturelle pour les problèmes unidimensionnels considérés, mais elle n'a pas de sens pour ∂D non régulière. Dans le dernier cas, les quantités v_i de (4.3) ne sont pas suffisamment régulières pour définir par récurrence toutes les fonctions v_k, $k < i$, nécessaires.

On abandonne donc, au voisinage de la frontière, les schémas aux différences simples et on en choisit d'autres, plus complexes, qui conduisent d'habitude à des équations à de nombreux coeffi-

cients non nuls. Cette tactique a été pour la première fois employée par Volkov [IV; 4]. Elle garantit une erreur de discrétisation qui ne varie pas lorsqu'on passe de l'intérieur du domaine dans une bande limitrophe. S'agissant de la méthode des éléments finis, la condition de continuité jusqu'à la frontière des coefficients du développement de l'erreur de discrétisation amène à choisir dans la bande limitrophe des fonctions de base plus complexes qui déterminent un nombre plus important d'inconnues des équations correspondantes (voir p. ex. Strang, Fix [III]).

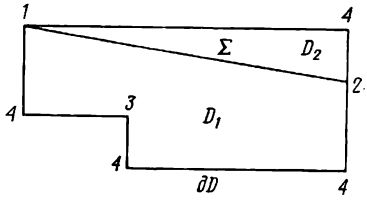


Fig. 19

Si la frontière ∂D est formée de segments de droites et il est possible de construire de façon simple et régulière jusqu'à la frontière le réseau et les éléments finis, la méthode d'extrapolation est entravée du fait que les dérivées de la solution sont indéfiniment croissantes au voisinage des points anguleux du domaine. Dans certains problèmes, cette croissance est telle qu'on n'obtient aucune précision, fût-elle minimale. Il en est de même de la régularité de la solution au voisinage de l'intersection de la frontière et des lignes de discontinuité de première espèce des coefficients de l'équation. La fig. 19 montre quatre types de points singuliers dont les trois premiers peuvent interdire la précision ordinaire de la solution (D_1, D_2 — domaines à l'intérieur desquels les coefficients sont réguliers; ∂D — frontière de la réunion de D_1 et D_2; Σ — ligne de discontinuité des coefficients).

Pour aplanir ces difficultés dans la méthode de Galerkin, on construit les équations aux différences en remplaçant les produits scalaires usuels par ceux pondérés avec les poids qui tendent vers 0 à mesure qu'on s'approche des points singuliers. Si la singularité coïncide par exemple avec l'origine des coordonnées, on utilise au lieu de la norme

$$\|u\|_{\mathring{W}_2^1} = \left[\int_\Omega \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] d\Omega\right]^{1/2} \tag{4.4}$$

naturelle au problème (4.1), (4.2) et aux fonctions s'annulant sur la frontière, la norme

$$\|u\|_{\alpha,\, \mathring{W}_2^1} = \left[\int_\Omega \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] (x^2+y^2)^\alpha \, d\Omega\right]^{1/2}. \tag{4.5}$$

Ici α est le degré de singularité de la solution qui est supposé connu à priori. On raisonne comme on le fait d'ordinaire dans le cas de Galerkin et on aboutit à une solution approchée u^h vérifiant la

majoration

$$\| u - u^h \|_{\alpha, \mathring{W}_2^1} \leqslant C \| u - \tilde{u} \|_{\alpha, \mathring{W}_2^1}, \tag{4.6}$$

$\tilde{u}$ étant l'interpolant de u dans l'espace des fonctions de base.

Il y a lieu de noter que cette méthode n'est pas universelle. En effet, on démontre (4.6) moyennant certaines propriétés de la solution. Or, ces propriétés ne sont pas des fois justes au voisinage des points singuliers.

Une autre voie à suivre est le resserrement du réseau ou des éléments finis pour que l'erreur de discrétisation devienne acceptable. Cette erreur est composée d'intégrales étendues sur les domaines élémentaires qui s'évaluent par

$$C h_i^{\beta} \| u \|_{k, i}, \tag{4.7}$$

où h_i est le diamètre du i-ième domaine élémentaire, $\| u \|_{k, i}$ la norme sur ce domaine, qui renferme la k-ième dérivée et C une constante indépendante de k et i. Une condition raisonnable à imposer

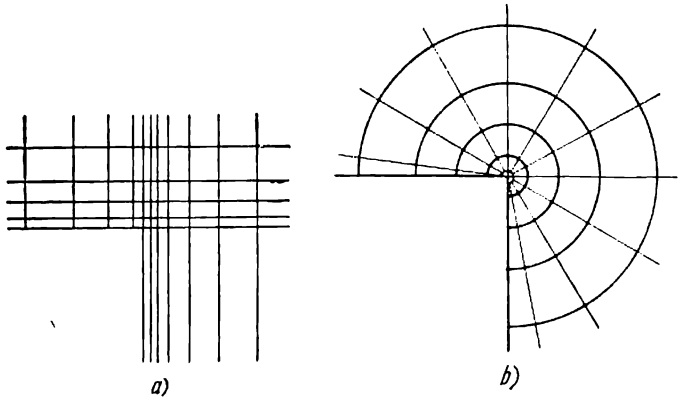


Fig. 20

à (4.7) est alors qu'ils coïncident dans D. A cet effet, on réduit les pas h_i en raison inverse des $\| u \|_{k,i}$ au voisinage des points singuliers. On ne saurait cependant resserrer les réseaux outre mesure parce que le nombre de conditionnement des systèmes algébriques des méthodes de Ritz et de Galerkin dépend essentiellement de $h_{\max}/h_{\min}$ ($h_{\min} = \min_i h_i$) qui augmente avec le passage aux maillages plus fins.

On voit sur la fig. 20 comment on resserre d'ordinaire les réseaux à proximité des points singuliers. Le cas *a*) s'avère en général plus simple sur le plan algorithmique que le cas *b*) car il n'alourdit guère

le programme tandis que le système polaire du cas *b*) est « recollé » au système rectangulaire (ou à d'autres réseaux polaires si les singularités sont 2 ou plus) (voir Volkov [IV; 4]). Le procédé *b*) coûte par contre moins du point de vue du nombre de nœuds. En effet, le resserrement n'a lieu qu'au voisinage d'un point singulier (dans le cas *a*) on resserre les nœuds suivant les lignes, y compris en dehors de cette région).

Voyons un procédé dû à Strang et Fix [III] (le lecteur peut en prendre connaissance dans Oganessian, Rivkind, Roukhovets [V; 8]). Il consiste à adjoindre aux fonctions de base usuelles du Chapitre 2 des fonctions complémentaires qui gouvernent très bien les dérivées de la solution au voisinage du point singulier. En effet, la solution u du problème (4.1), (4.2) avec le second membre dans $W_2^k(D)$ est représentable par la somme d'une fonction $v \in W_2^{k+2}(D)$ régulière et d'un nombre fini de termes qui s'écrivent en coordonnées polaires centrées au point singulier: $w_i = \mu_i(\varphi)\, r^{\gamma_i} \ln^{p_i} r$, $\mu(\varphi)$ étant une fonction analytique de la variable angulaire φ, γ_i une constante réelle positive et p_i un entier non négatif. Les fonctions de base de cette forme spéciale ont été obtenues par Nikolski [I; 5]. Dans chaque cas concret, elles sont en général définies à un facteur constant près. Ainsi, on cherche dans la méthode de Galerkin non seulement les poids du développement de la composante régulière suivant les fonctions de base de la méthode des éléments finis, mais aussi les poids des fonctions spéciales. L'inconvénient principal de cette méthode provient des fonctions w_i qui perturbent l'arrangement des éléments non nuls du système algébrique parce que leurs supports empiètent sur de nombreux domaines élémentaires.

Avec les fonctions de base spéciales qui décrivent très bien le comportement de la solution au voisinage des points singuliers, on utilise à l'intérieur du domaine des éléments simples et on atteint une grande précision en extrapolant sur le pas du réseau. On va se borner, pour le problème (4.1), (4.2), au schéma général de l'algorithme et nous le justifierons dans ses grandes lignes car le cas en dimension deux se prête à de nombreuses variantes absentes de ce livre. On note que le problème d'augmenter la précision utilise une information riche sur la régularité des solutions des équations elliptiques dans des domaines de diverses formes (pour cette information, consulter Volkov [IV; 4], Lébédev [IV; 15]).

Nous avons supposé, pour fixer les idées, que le domaine est un rectangle de côtés de longueur a et b. On établit, à l'aide de deux familles de droites parallèles contenant a et b, un réseau rectangulaire régulier de pas $h_x = a/N$ et $h_y = b/N$, N étant un entier >2. On effectue la triangulation en menant dans chaque rectangle une diagonale montant de gauche à droite. Il y a intérêt à introduire le paramètre $h = 1/N$ caractéristique de la dimension du réseau et relié à h_x, h_y par les égalités $h_x = ah_1$, $h_y = bh_1$.

On fait correspondre à chaque nœud intérieur X_i une fonction de base φ_i (Chap. 2) linéaire sur chaque triangle. On introduit, en plus de φ_i en nombre $(N-1)^2$, les fonctions de base spéciales ψ_k égales à w_i dans une région autour des points singuliers (points anguleux du rectangle et points où la frontière coupe les lignes de discontinuité éventuelles des coefficients qui se confondent avec les lignes du réseau de discrétisation) et prolongées régulièrement sur le domaine restant de façon qu'on respecte la condition aux limites (4.2). Le nombre de ces fonctions spéciales varie d'un problème à l'autre selon la précision désirée.

En procédant par Ritz ou par Galerkin on trouve la solution

$$u^h(X) = \sum_{i=1}^{(N-1)^2} \alpha_i^h \varphi_i(X) + \sum_{k=1}^{M} \beta_k^h \psi_k(X) \tag{4.8}$$

qui admet aux nœuds du réseau (aux sommets des triangles) un développement suivant h, à savoir

$$u^h(X) = u(X) + \sum_{l=1}^{m-1} h^{2l} v_l(X) + h^{2m} \xi_h(X), \tag{4.9}$$

où $v_l(X)$ sont indépendantes de l et $\xi_h(X)$ est une fonction discrète bornée. Une fois le développement justifié, on supprime les termes en $h^2, \ldots, h^{2m-2}$ et on obtient $u(x)$ en $O(h^{2m})$.

On se rappelle, pour justifier (4.9), que les poids α_i^h et β_k^h de la formule (4.8) se développent également suivant les puissances de h :

$$\begin{aligned} \alpha_i^h &= v(X_i) + \sum_{l-1}^{m-1} h^{2l} Z_l(X_i) + h^{2m} \eta_h(X_i), \\ \beta_k^h &= \gamma_k + \sum_{l=1}^{m-1} h^{2l} \delta_{k,l} + h^{2m} \zeta_{k,h}, \end{aligned} \tag{4.10}$$

avec $v(X)$ la composante régulière de $u(X)$; $Z_l(X)$ des fonctions indépendantes de h ; η_h une fonction discrète bornée pour $h \to 0$ et γ_k, $\delta_{k,l}$, $\zeta_{k,h}$ des nombres tels que γ_k et $\delta_{k,l}$ ne dépendent pas de h et $\zeta_{k,h}$ soient bornés quand $h \to 0$.

On part du développement (4.10) et on formule des conditions (problèmes différentiels auxiliaires) qui définissent $Z_l(X)$ et les coefficients constants $\delta_{k,l}$ des fonctions $\varphi_k(X)$ dans la décomposition des solutions des problèmes auxiliaires en les composantes régulières $v_l(X)$ et les fonctions de base spéciales $\psi_k(X)$. Le schéma de mise en œuvre suggère au plus le mode opératoire qui doit être justifié de façon exhaustive dans chaque cas concret.

CHAPITRE 7

PROBLÈMES INVERSES (POSITION ET MÉTHODES NUMÉRIQUES)

On qualifie d'inverses divers types de problèmes de la physique mathématique. De vastes classes de problèmes inverses sont étudiées par exemple par Lévitan [XVI; 8], Lavrentiev, Romanov, Vasiliev [XVI].

Nous nous occuperons de deux sortes de problèmes inverses: 1) définir l'état d'un processus à un temps antérieur (on demande par exemple la répartition de température à l'instant initial connaissant le champ thermique à une époque donnée); 2) reconstituer un opérateur de forme connue à coefficients à déterminer par une information sur les fonctionnelles des solutions (le problème inverse pour l'équation de Sturm-Liouville consiste par exemple à définir le coefficient dans l'équation différentielle du second ordre connaissant les propriétés de la fonction spectrale d'un problème aux limites).

De nombreux problèmes inverses de la physique mathématique s'avèrent mal posés au sens classique: aux faibles variations des fonctionnelles il correspond de grandes variations des solutions. Le concept de problème bien posé a été introduit par Hadamard qui a également donné un exemple de problème mal posé (problème de Cauchy pour l'équation de Laplace).

L'intérêt des problèmes mal posés a été méconnu jusqu'au jour où la nécessité s'est fait sentir d'interpréter les données géophysiques. Une grande contribution aux méthodes de résolution et à la théorie des problèmes qui ne sont pas bien posés au sens classique (ou selon Hadamard) a été faite par Tikhonov [XVI; 17], M.M. Lavrentiev [II] [XVI; 5, 25], John [XVI], Ivanov [XVI; 1].

On sait que la solution des problèmes mal posés au sens classique devient stable par rapport aux variations des données si l'on impose des contraintes supplémentaires à l'ensemble des solutions admissibles. Ces problèmes sont donc par définition *conditionnellement bien posés.*

On a introduit la notion de famille régularisante: on associe au problème conditionnellement bien posé une famille de problèmes bien posés au sens classique (*famille régularisante*) qui dépend d'un

paramètre et telle que la suite de solutions converge vers la solution du problème proposé lorsque le paramètre tend vers une limite. On a montré qu'avec un paramètre de régularisation convenable dépendant de la précision des données, la solution associée aux données approchées d'un problème de la famille est solution approchée du problème conditionnellement bien posé.

Les auteurs cités ci-dessus ont étudié une vaste gamme de problèmes conditionnellement bien posés. Dans ce chapitre nous nous intéresserons à une seule classe de problèmes inverses liés à des équations d'évolution et nous les aborderons par une méthode de régularisation procédant du développement en série de Fourier.

7.1. Préliminaires

Soit F un espace de Hilbert des fonctions f dont chacune se développe en série de Fourier suivant un système biorthogonal complet de fonctions $\{u_n\}$ et $\{u_n^*\}$. Ainsi

$$f = \sum_{n=1}^{\infty} f_n u_n, \tag{1.1}$$

où

$$f_n = (f,\ u_n^*). \tag{1.2}$$

Formons un sous-espace Φ de F de toutes les f dont les développements (1.1) possèdent au plus N composantes non nulles associées aux plus grandes perturbations:

$$f = \sum_{n=1}^{N} f_n u_n. \tag{1.3}$$

Ce sous-espace est évidemment inclus dans l'espace des fonctions qui se développent en série de Fourier.

Considérons le problème

$$A\varphi = f, \tag{1.4}$$

A étant un opérateur linéaire possédant un inverse non borné. On aborde (1.4) par la méthode itérative

$$\varphi^{j+1} = \varphi^j - \tau\,(A\varphi^j - f), \qquad \varphi^0 = 0, \tag{1.5}$$

ou

$$\varphi^{j+1} = T\varphi^j + \tau f, \qquad \varphi^0 = 0, \tag{1.6}$$

avec $T = E - \tau A$ l'opérateur de transition.

On suppose que le problème (1.4) admet une solution et une seule qui appartient à Φ et qu'on met sur F tout entier la norme

$$\|T\|_F^2 = \sup_{\varphi \in F} \frac{(T\varphi,\ T\varphi)}{(\varphi,\ \varphi)} = \beta\,(T^*T) > 1, \qquad \beta\,(T) > 1,$$

et sur Φ

$$\| T \|_{\Phi}^2 = \sup_{\varphi \in \Phi} \frac{(T\varphi,\ T\varphi)}{(\varphi,\ \varphi)} < 1. \tag{1.7}$$

Avec les hypothèses faites, le processus itératif (1.6) converge sur $\varphi^j \in \Phi$ et diverge sur les éléments de F tout entier. Pour qu'il y ait convergence, il faut donc que la solution approchée φ^j se trouve dans Φ à chaque pas d'itération. Cela ne présente pas de difficulté sur le plan constructif.

Posons $\varphi^{j-1} \in \Phi$ à un pas et utilisons la relation de récurrence (1.6), il vient un nouveau élément φ^j. La fonction $T\varphi^{j-1}$ peut ne pas être dans Φ. Pour que $\varphi^j \in \Phi$, on développe en série de Fourier

$$\varphi^j = \sum_{n=1}^{\infty} \varphi_n^j u_n$$

et on rejette tous les termes d'indice $n > N$. Algorithmiquement, il y a intérêt à définir N premiers coefficients de Fourier $\varphi_k^j = (\varphi^j,\ u_k^*)$, $k = 1, 2, \ldots, N$, et à former la somme finie

$$\varphi^j = \sum_{k=1}^{N} \varphi_k^j u_k. \tag{1.8}$$

Si l'on continue par itérations, toutes les fonctions φ^j, solutions approchées du problème, appartiennent à Φ. Si l'on suppose par exemple l'orthogonalité de la base $\{u_n\}$, la suite $\{\varphi^j\}$ converge vers une fonction φ^∞ qu'on prend pour solution approchée de l'équation (1.4).

Nous avons considéré un algorithme simple d'approximations internes. On pourrait en décrire plusieurs autres.

Soit (1.4) à A positif symétrique et supposons que $f \in \Phi$, $\varphi \in \Phi$, Φ étant un sous-espace de l'espace hilbertien F.

Nous dirons que le problème (1.4) est conditionnellement bien posé si sa solution appartient à un $\Phi \subset F$ et on a l'estimation à priori

$$\| \varphi \|_{\Phi} \leqslant M_0 \| f \|_{\Phi}, \tag{1.9}$$

avec M_0 une constante et

$$\| f \|_{\Phi} < \infty. \tag{1.10}$$

On résout (1.4) par rapport à φ:

$$\varphi = \mathrm{A}^{-1} f \tag{1.11}$$

et on évalue φ en norme sur ce sous-espace, il vient

$$\| \varphi \|_{\Phi} \leqslant \| \mathrm{A}^{-1} \|_{\Phi} \| f \|_{\Phi}. \tag{1.12}$$

Il en résulte que pour que le problème soit bien posé il suffit d'être dans la condition à priori

$$\| A^{-1} \|_{\Phi} \leqslant M. \tag{1.13}$$

Chose importante, si A est un opérateur symétrique, alors

$$\| A^{-1} |'_{\Phi} = \sup_{\varphi \in \Phi} \frac{(A^{-1}\varphi, \varphi)}{(\varphi, \varphi)}, \qquad \| A^{-1} \|_F = \sup_{\varphi \in F} \frac{(A^{-1}\varphi, \varphi)}{(\varphi, \varphi)}. \tag{1.14}$$

On vérifie sans peine que si $\Phi \subset F$, alors

$$\| A^{-1} \|_{\Phi} \leqslant \| A^{-1} \|_F = \frac{1}{\alpha(A)}.$$

Ainsi, la propriété, pour le problème, d'être bien posé est déterminée, en plus de (1.9), par la condition équivalente (1.13). Il y a plus. Si la constante M est donnée, le problème crucial que pose l'organisation du calcul, est de définir le sous-espace Φ qui renferme, en plus de la solution φ, la suite de ses approximations φ^j.

Un autre aspect de la résolution des problèmes conditionnellement bien posés est la précision des données initiales. Dans ces problèmes, on voit intervenir au moins une erreur due à l'approximation de (1.4) par un problème aux différences ou à l'imprécision des coefficients de A et de la fonction f.

Soit $\bar{A}$, $\bar{f}$ et $\bar{\varphi}$ l'opérateur exact, le vecteur exact et la solution exacte respectivement du problème

$$\bar{A}\bar{\varphi} = \bar{f},$$

et

$$A = \bar{A} + \delta A, \qquad f = \bar{f} + \delta f. \tag{1.15}$$

On suppose connues les erreurs à priori

$$\| \delta A \|_{\Phi} \leqslant \varepsilon_1, \tag{1.16}$$

$$\| \delta f \|_{\Phi} \leqslant \varepsilon_2. \tag{1.17}$$

Résoudre le problème (1.4) revient à former une suite d'itérations qui génèrent de nouvelles approximations dans Φ. Si A est une matrice définie positive, les problèmes se prêtent à tout un jeu de procédés itératifs convergents vers la solution de (1.4) (voir Chap. 4). Ceci étant, le sous-espace Φ des solutions approchées est identique à l'espace hilbertien (euclidien) F tout entier. Il s'agit là d'une particularité agréable des problèmes à matrices positives. Le processus itératif

$$\varphi^{j+1} = \varphi^j - \tau (A\varphi^j - f), \qquad \varphi^0 = 0 \tag{1.18}$$

converge pour une bonne valeur du paramètre τ et on l'optimise par exemple, compte tenu de l'information à priori (1.16), (1.17), en choisissant le nombre j_0 de pas d'itération (voir n° 4.6).

L'opérateur A du problème (1.4) est supposé symétrique à spectre ayant une partie positive et une partie négative.

On montre que sous ces hypothèses il y a divergence du processus (1.18).

En effet, soit

$$\varphi = \sum_n \varphi_n u_n, \qquad f = \sum_n f_n u_n \tag{1.19}$$

et $\{u_n\}$ un système orthonormé complet de fonctions propres de l'opérateur A. Portons (1.19) dans (1.18) et multiplions scalairement par u_n, il vient les relations de récurrence pour les coefficients de Fourier

$$\varphi_n^{j+1} = \varphi_n^j - \tau(\lambda_n \varphi_n^j - f_n), \qquad \varphi_n^0 = 0$$

ou pour le résidu $\xi^j = A\varphi^j - f$:

$$\xi_n^{j+1} = (1 - \tau\lambda_n)\,\xi_n^j, \qquad \xi_n^0 = -f_n. \tag{1.20}$$

On résout (1.20):

$$\xi_n^j = -(1 - \tau\lambda_n)^j f_n. \tag{1.21}$$

Donc

$$\xi^j = -\sum_n (1 - \tau\lambda_n)^j f_n u_n. \tag{1.22}$$

Les itérations (1.18) ne convergent naturellement que si

$$\lim_{j\to\infty} \xi^j = 0.$$

Si l'opérateur A n'a d'autres valeurs propres que les nombres positifs de l'intervalle

$$\alpha(A) \leqslant \lambda_n(A) \leqslant \beta(A),$$

le processus (1.18) devient convergent à condition de choisir $\tau > 0$ dans les limites

$$0 < \tau < 2/\beta. \tag{1.23}$$

Mais la matrice symétrique A possède par hypothèse des valeurs propres de deux signes. Etant donné τ de l'intervalle (1.23), toutes les composantes du résidu qui correspondent à λ positifs, diminuent d'une itération à l'autre comme T_n^j, où $T_n = (1 - \tau\lambda_n)^j < 1$ et j est l'exposant. Quant aux composantes associées à λ négatifs, on a

$$T_n^j = (1 - \tau\lambda_n)^j > 1,$$

et ces composantes augmentent, d'où la divergence mentionnée.

Ainsi, le processus itératif (1.18) avec les fonctions d'essai φ^j appartenant à l'espace de Hilbert tout entier, possède la propriété de divergence.

Un exemple d'approximations internes pour Φ est la méthode du plus petit résidu à trois niveaux (voir point 4.3.2)

$$\varphi^{j+1} = \varphi^j - \tau^j (A\varphi^j - f) - \gamma_j A^* (A\varphi^j - f).$$

Quelques mots enfin sur la pratique de la résolution des problèmes conditionnellement bien posés. On les ramène d'ordinaire à des systèmes d'équations linéaires à matrices mal conditionnées de forme générale. On recourt de règle à la technique du plus petit résidu à n niveaux (voir point 4.3.2) qui garantit une bonne convergence. On applique de même la méthode du gradient conjugué aux équations symétrisées par la transformation de Gauss. La dernière tactique sera examinée à propos des problèmes d'évolution inverses (voir n° 7.3). S'agissant des méthodes itératives, l'algorithme s'arrête dès que la norme du résidu devient approximativement égale à l'erreur à priori sur les données, i.e. $\| \xi^j \| \approx \varepsilon_1 + \varepsilon_2$. Etant données les erreurs à priori, on obtient alors, comme on l'a dit au n° 4.6, une solution exacte au possible.

7.2. Problèmes d'évolution inverses et méthode de Fourier

Soit A une matrice définie positive indépendante du temps, à spectre réel compris dans l'intervalle $\alpha(A) \leqslant \lambda \leqslant \beta(A)$ et φ une fonction vectorielle solution du problème de Cauchy

$$\begin{gathered} \frac{d\varphi}{dt} - A\varphi = 0, \qquad 0 \leqslant t \leqslant t_0, \\ \varphi = g \quad \text{pour} \quad t = 0, \end{gathered} \tag{2.1}$$

g étant une valeur donnée de φ à l'instant initial.

Considérons deux problèmes spectraux

$$Au = \lambda u, \qquad A^* u^* = \lambda u^* \tag{2.2}$$

et supposons qu'ils définissent deux bases biorthogonales de fonctions propres $\{u_n\}$ et $\{u_n^*\}$. Les fonctions φ et g sont mises sous forme de sommes de Fourier

$$\varphi = \sum_n \varphi_n u_n, \qquad g = \sum_n g_n u_n. \tag{2.3}$$

Portons dans (2.1) et multiplions scalairement par u_n^*, il vient le système d'équations différentielles ordinaires pour les coefficients de Fourier

$$\begin{gathered} \frac{d\varphi_n}{dt} - \lambda_n \varphi_n = 0, \\ \varphi_n = g_n \quad \text{pour} \quad t = 0, \\ n = 1, 2, \ldots, N. \end{gathered} \tag{2.4}$$

La solution de chaque équation de ce système est de la forme

$$\varphi_n = g_n e^{\lambda_n t}, \qquad n = 1, 2, \ldots, N, \tag{2.5}$$

si bien que la solution du problème (2.1) admet la représentation

$$\varphi(t) = \sum_{n=1}^{N} g_n e^{\lambda_n t} u_n. \tag{2.6}$$

Ainsi, on a établi que la solution du problème (2.1) est représentée par la somme de Fourier dont chaque terme croît de façon exponentielle dans le temps en fonction de la n-ième valeur propre λ_n.

Si l'on veut connaître une solution physiquement définie au problème dans l'intervalle $0 \leqslant t \leqslant t_0$, on considère l'analogue bien posé de (2.1)

$$\begin{gathered} \frac{d\varphi}{dt} - A\varphi = 0, \qquad 0 \leqslant t \leqslant t_0, \\ \varphi = h \quad \text{pour} \quad t = t_0. \end{gathered} \tag{2.7}$$

On procède comme plus haut et on obtient

$$\varphi = \sum_{n=1}^{N} h_n e^{-\lambda_n (t_0 - t)} u_n. \tag{2.8}$$

On astreint la solution (2.8) à coïncider pour $t = 0$ avec le vecteur g du problème (2.1) et on établit la relation entre les coefficients de Fourier de la fonction h et ceux de g:

$$g_n = h_n e^{-\lambda_n t_0}. \tag{2.9}$$

Ainsi, on reconstitue sans peine g par h:

$$g = \sum_{n=1}^{N} h_n e^{-\lambda_n t_0} u_n, \tag{2.10}$$

et il correspond à de petites erreurs sur h (ou sur h_n) des erreurs faibles sur g. Mais il s'agit d'un problème inverse du problème proposé car on doit reconstituer h par la formule

$$h = \sum_{n=1}^{N} g_n e^{\lambda_n t_0} u_n \tag{2.11}$$

à l'aide d'une certaine information sur g.

S'il s'agissait d'une information exacte et si les calculs portaient sur un nombre infini de chiffres significatifs, la formule (2.11) aurait donné h de façon simple. Or, la fonction g est entachée d'une erreur connue à priori et on travaille sur des caractères en nombre limité, d'où les erreurs d'arrondi. La reconstitution de h par (2.11) n'est donc pas si facile qu'on croit.

Supposons que le chercheur désireux de traiter les données expérimentales à la lumière du problème d'évolution (2.1) connaît à l'avance le système de fonctions propres u_n et que s'il développe les données initiales (les fonctions g) suivant ce système, il en tire une information utile et évalue assez exactement l'erreur sur chaque composante de Fourier g_n.

Dans le cas de problèmes répétitifs de caractère statistique, les méthodes de corrélation bien élaborées permettent d'augmenter de façon sensible la précision sur g_n même si l'erreur à une seule mesure dépasse considérablement l'information utile. Toujours est-il que le traitement préalable des résultats d'observation aide à se prononcer sur la grandeur de l'erreur systématique (de l'erreur aléatoire dans le cas d'une mesure isolée) sur g_n. On a donc pour tout n

$$g_n = \bar{g}_n (1 + \delta_n),$$

où $\bar{g}_n$ est la valeur exacte (inconnue à priori!) et δ_n l'erreur relative qu'on supposera connue.

L'erreur δ_n s'avère d'ordinaire minimale pour les perturbations de plus grandes longueurs d'onde et elle croît vite dans le sens des harmoniques supérieurs qui décrivent en général les singularités « petites » de la solution. Aussi les coefficients g_n gouvernent principalement à partir d'un numéro l'erreur de données initiales. Selon la formule (2.11), ce sont les composantes haute fréquence qui croissent exponentiellement le plus. On doit donc écraser ces harmoniques parasites sinon on risque d'obtenir les résultats fantaisistes. En effet, g_n de ces harmoniques ne renferment en fait aucune information utile. Si on les multiplie par $e^{\lambda_n t_0}$ importants, ils interviennent avec un grand poids dans h et risquent d'altérer catastrophiquement la solution. Notre préoccupation majeure est donc de déterminer le contenu informationnel des g_n.

Supposons que, d'après l'information à priori, l'erreur relative affectant n_0 premiers g_n est inférieure à η, i.e. $\delta_n < \eta$, η étant la plus grande erreur admissible. L'algorithme de reconstitution (2.11) est alors un analogue de la procédure de génération des éléments de Φ du problème (2.1) à condition d'éliminer les harmoniques parasites de la série (2.11). Finalement,

$$h = \sum_{n=1}^{n_0} g_n e^{\lambda_n t_0} u_n. \tag{2.12}$$

L'algorithme de résolution du problème spectral particulier se base sur une méthode itérative du numéro 1.1. S'il faut former un ensemble des premières (plus « grandes ») fonctions propres u_n ou u_n^* et l'ensemble des valeurs propres associées, on utilise l'algorithme d'orthogonalisation du même numéro.

7.3. Problème d'évolution inverse à opérateur dépendant du temps

Considérons le problème d'évolution

$$\begin{gathered}\frac{d\varphi}{dt} - A(t)\varphi = 0, \qquad 0 \leqslant t \leqslant t_0, \\ \varphi = g \quad \text{pour} \quad t = 0,\end{gathered} \tag{3.1}$$

l'opérateur $A > 0$ étant fonction du temps. On suppose une fois de plus que (3.1) intervient dans la réduction d'un problème de la physique mathématique par rapport aux variables d'espace à un système d'équations différentielles ordinaires. Dans ce cas la méthode de Fourier tombe en défaut et on utilise des procédés numériques.

Voyons l'un des algorithmes possibles. On associe au problème (3.1) un problème modèle voisin dans un sens, à savoir

$$\begin{gathered}\frac{d\bar{\varphi}}{dt} - \bar{A}\bar{\varphi} = 0, \qquad 0 \leqslant t \leqslant t_0, \\ \bar{\varphi} = g \quad \text{pour} \quad t = 0,\end{gathered} \tag{3.2}$$

l'opérateur $\bar{A} > 0$ étant indépendant du temps, à spectre positif $\alpha(\bar{A}) \leqslant \lambda(\bar{A}) \leqslant \beta(\bar{A})$ et voisin de $A(t)$ dans un sens. On suppose, pour fixer les idées,

$$A(t) = \bar{A} + \delta A(t), \tag{3.3}$$

avec

$$\| \delta A(t) \| \ll \| \bar{A} \| \tag{3.4}$$

quel que soit t de l'intervalle $0 \leqslant t \leqslant t_0$.

Le problème (3.2) nous fournira une information à priori nécessaire pour organiser le calcul de la solution de (3.1). On définit, par des techniques du n° 7.2 et du n° 1.1, m fonctions propres u_n, u_n^* porteuses d'information (du point de vue des erreurs de données) et les valeurs propres associées λ_n, $n = 1, 2, \ldots, m$. On supprime les composantes restantes de la série de Fourier pour g_n, $n = m + 1$, $m + 2, \ldots, N$, parce que les erreurs commises en définissant ces coefficients dépassent (de façon substantielle des fois) l'information utile. On obtient

$$\bar{g} = \sum_{n=1}^{m} g_n u_n, \tag{3.5}$$

où

$$g_n = (g, u_n^*),$$

et la solution $\bar{\varphi}$ du problème modèle sur l'intervalle $0 \leqslant t \leqslant t_0$ s'écrit

$$\bar{\varphi}(t) = \sum_{n=1}^{m} g_n e^{\lambda_n t} u_n. \tag{3.6}$$

Passons à la résolution numérique de ce problème et utilisons par exemple le schéma aux différences exact à l'ordre 2 en $\Delta t = \tau$:

$$\frac{\bar{\varphi}^{j+1} - \bar{\varphi}^j}{\tau} - \bar{A}\,\frac{\bar{\varphi}^{j+1} + \bar{\varphi}^j}{2} = 0, \qquad \bar{\varphi}^0 = g, \qquad j = 1,\ 2,\ \ldots,\ j_0. \tag{3.7}$$

On aborde (3.7) par la méthode de Fourier et on suppose connu tout l'ensemble des fonctions propres u_n et u_n^*, hypothèse dont le seul but est de faciliter l'analyse théorique et de fournir unè information à priori sur le comportement de la solution. On a

$$\bar{\varphi}^j = \sum_{n=1}^{N} \bar{\varphi}_n^j u_n, \tag{3.8}$$

et on aboutit moyennant (3.7) aux récurrences

$$\bar{\varphi}_n^{j+1} = \frac{1 + \frac{\tau\lambda_n}{2}}{1 - \frac{\tau\lambda_n}{2}}\,\bar{\varphi}_n^j, \qquad \bar{\varphi}_n^0 = g_n, \qquad j = 1,\ 2,\ \ldots,\ j_0. \tag{3.9}$$

Donc

$$\bar{\varphi}_n^j = \left[\frac{1 + \frac{\tau\lambda_n}{2}}{1 - \frac{\tau\lambda_n}{2}}\right]^j g_n. \tag{3.10}$$

Ainsi

$$\bar{\varphi}^j = \sum_{n=1}^{N} T_n^j g_n u_n, \tag{3.11}$$

où

$$T_n = \frac{1 + \frac{\tau\lambda_n}{2}}{1 - \frac{\tau\lambda_n}{2}}.$$

On fait l'hypothèse suivante : le pas τ est choisi à partir de la condition de non-nullité du dénominateur pour tout n. On a

$$\tau < \frac{2}{\beta(\bar{A})}. \tag{3.12}$$

Il y a lieu de noter que cette condition est compatible avec la condition d'approximation des perturbations à grande échelle.

Une analyse formelle de la solution (3.11) du problème modèle montre que tous les $T_n > 1$ et que les amplitudes des harmoniques haute fréquence associés à n élevés croissent vite avec le numéro. Par conséquent, on a pour ces harmoniques $T_n \gg 1$ et à fortiori $T_n^j \gg 1$. Puisque le traitement des données initiales g a eu lieu en supprimant toutes les composantes de Fourier de (3.5) à partir de $n = m + 1$, cela suffit visiblement pour que la somme

$$\bar{g} = \sum_{n=1}^{m} g_n u_n$$

donne une solution ayant le même nombre de termes

$$\bar{\varphi}^j = \sum_{n=1}^{m} T_n^j g_n u_n. \tag{3.13}$$

Il en serait ainsi si le calcul sur ordinateur se faisait avec un nombre infini de chiffres significatifs. Or, les erreurs d'arrondi résultant du caractère borné du mot de machine font apparaître dès le début des calculs les composantes g_n, $n > m$. Bien que petites, ces composantes constituent une partie importante proportionnelle à $T_n^j \gg 1$ de la solution. Ces erreurs peuvent altérer de façon sensible la solution. On évite une croissance catastrophique des erreurs sur les composantes de Fourier à haute fréquence si l'on trouve un artifice pour transformer automatiquement tout élément de l'espace vectoriel F en l'élément d'un sous-espace Φ.

On définit Φ comme étant l'ensemble des éléments tels qu'au cours du calcul les amplitudes des $N - m$ derniers harmoniques de leur série de Fourier procédant suivant le système de fonctions propres u_n augmentent à chaque pas au plus comme plusieurs amplitudes du dernier harmonique informationnel d'indice m. Dans ce sous-espace, les erreurs d'arrondi sur les éléments croissent au plus comme l'amplitude de la m-ième composante de Fourier, ce qui garantit la propriété du schéma de calcul d'être bien posé.

M.M. Lavrentiev [II], Lattès et Lions [XVI] ont proposé de remplacer $\bar{A}$ du problème modèle (3.2) par l'opérateur $\bar{A}_\varepsilon = \bar{A} - \varepsilon \bar{A}^2$, auquel cas (3.2) fait place à

$$\begin{gathered} \frac{d\bar{\varphi}_\varepsilon}{dt} - \bar{A}\bar{\varphi}_\varepsilon = -\varepsilon \bar{A}^2 \bar{\varphi}_\varepsilon, \qquad 0 \leqslant t \leqslant t_0, \\ \bar{\varphi}_\varepsilon = g \quad \text{pour} \quad t = 0, \end{gathered} \tag{3.14}$$

où ε est pour l'instant un paramètre quelconque. On choisit ε sous la condition que la solution du problème reste dans Φ. Supposons, pour simplifier les raisonnements, que $\bar{A} = \bar{A}^*$ et soit le schéma

aux différences

$$\frac{\bar{\varphi}_\varepsilon^{j+1}-\bar{\varphi}_\varepsilon^{j}}{\tau}-(\bar{A}-\varepsilon\bar{A}^2)\frac{\bar{\varphi}_\varepsilon^{j+1}+\bar{\varphi}_\varepsilon^{j}}{2}=0, \qquad \bar{\varphi}_\varepsilon^{0}=g. \tag{3.15}$$

On cherche la solution de (3.15) en développant en série de Fourier suivant les fonctions propres de $\bar{A}$. On a

$$\bar{\varphi}_\varepsilon^{j}=\sum_{n=1}^{N}\left[\frac{1+\frac{\tau\lambda_n}{2}-\varepsilon\frac{\tau\lambda_n^2}{2}}{1-\frac{\tau\lambda_n}{2}+\varepsilon\frac{\tau\lambda_n^2}{2}}\right]^j g_n u_n. \tag{3.16}$$

On astreint ε à la condition que l'erreur relative commise sur le m-ième harmonique par l'introduction de l'opérateur $\varepsilon\bar{A}^2$ ne dépasse pas η (on prend d'ordinaire $\eta < 1$ selon les parts respectives de l'information utile et du « bruit » dans l'harmonique $n = m$). Cette condition donne

$$\eta\frac{\tau\lambda_m}{2}=\varepsilon\frac{\tau\lambda_m^2}{2}, \tag{3.17}$$

d'où

$$\varepsilon=\frac{\eta}{\lambda_m}. \tag{3.18}$$

Ainsi, on a défini une quantité à priori très importante pour le calcul. On voit aisément qu'avec ε de (3.18), les amplitudes de tous les harmoniques d'indice $n > m$ croissent avec le temps au plus comme T_m^j.

Nous aurons besoin d'une autre quantité à priori. Considérons

$$\bar{\varphi}^{j}=\sum_{n=1}^{m} g_n e^{\lambda_n t_j} u_n \tag{3.19}$$

et

$$\bar{\varphi}_\varepsilon^{j}=\sum_{n=1}^{N} g_n T_n^j(\varepsilon)\, u_n, \tag{3.20}$$

où

$$T_n(\varepsilon)=\frac{1+\frac{\tau\lambda_n}{2}-\varepsilon\frac{\tau\lambda_n^2}{2}}{1-\frac{\tau\lambda_n}{2}+\varepsilon\frac{\tau\lambda_n^2}{2}}.$$

Puisque la solution $\bar{\varphi}_\varepsilon^j$ se trouve dans Φ, on la remplace sans grande erreur par

$$\bar{\varphi}_\varepsilon^{j}=\sum_{n=1}^{m} g_n T_n^j(\varepsilon)\, u_n, \tag{3.21}$$

où on s'est borné aux m premiers termes de la série. On cherche la solution (3.21) de façon constructive à l'aide des systèmes obtenus de u_n et u_n^*, $n = 1, 2, \ldots, m$. On trouve les quantités $\bar{\varphi}^j$, $\bar{\varphi}_\varepsilon^j$, $j = 1, 2, \ldots, j_0$, moyennant (3.19) et (3.21), puis on introduit les vecteurs

$$\bar{\varphi} = \begin{Vmatrix} \bar{\varphi}^1 \\ \bar{\varphi}^2 \\ \cdots \\ \bar{\varphi}^{j_0} \end{Vmatrix}, \qquad \bar{\varphi}_\varepsilon = \begin{Vmatrix} \bar{\varphi}_\varepsilon^1 \\ \bar{\varphi}_\varepsilon^2 \\ \cdots \\ \bar{\varphi}_\varepsilon^{j_0} \end{Vmatrix}$$

et on calcule la norme

$$\| \bar{\varphi} - \bar{\varphi}_\varepsilon \| = \delta \tag{3.22}$$

qui est justement la dernière quantité à priori nécessaire (les deux autres sont définies par (3.12) et (3.18)).

Enonçons l'algorithme de résolution numérique du problème initial (3.1). Etant données les considérations ci-dessus, on construit l'approximation de celui-ci :

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} - (A_j - \varepsilon A_j^2) \frac{\varphi^{j+1} + \varphi^j}{2} = 0, \qquad \varphi^0 = g, \tag{3.23}$$

avec τ et ε dictés par l'analyse du modèle simple étudié à priori :

$$\tau < \frac{2}{\beta(\bar{A})}, \qquad \varepsilon = \frac{\eta}{\lambda_m(\bar{A})}. \tag{3.24}$$

On introduit les vecteurs

$$\varphi = \begin{Vmatrix} \varphi^1 \\ \varphi^2 \\ \cdots \\ \varphi^{j_0} \end{Vmatrix}, \qquad f = \begin{Vmatrix} -R_0 g \\ 0 \\ \cdots \\ 0 \end{Vmatrix}$$

et la matrice

$$\Lambda = \begin{Vmatrix} -S_0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ R_1 & -S_1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & R_2 & -S_2 & 0 & \ldots & 0 & 0 \\ 0 & 0 & R_3 & -S_3 & \ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \ldots & R_{j_0-1} & -S_{j_0-1} \end{Vmatrix},$$

où

$$S_j = E - \frac{\tau}{2}(A_j - \varepsilon A_j^2), \qquad R_j = E + \frac{\tau}{2}(A_j - \varepsilon A_j^2),$$

$$A_j = A(t_{j+1/2}).$$

Le problème correspondant s'énonce

$$\Lambda\varphi = f. \tag{3.25}$$

On le symétrise en multipliant par Λ^* :

$$\Lambda^*\Lambda\varphi = \Lambda^* f \tag{3.26}$$

et on formule un processus itératif. On utilise en particulier la méthode du gradient conjugué dans laquelle on se passe de connaître à priori les bornes du spectre de $\Lambda^*\Lambda$.

Les approximations successives n'épuisent pas l'algorithme. Il faut de plus définir le nombre optimal k_0 d'itérations qui donnent la plus grande précision pour les conditions à priori données. Ce nombre peut être obtenu avec une précision médiocre, si bien que l'estimation à priori (3.22) du problème modèle joue pour le problème initial (3.1). On suppose donc que

$$\| \varphi - \varphi_\varepsilon \| = \delta, \tag{3.27}$$

φ étant la solution exacte de (3.1) aux nœuds du réseau et φ_ε la solution du problème aux différences à opérateur régularisant. On continue naturellement le processus itératif tant que l'erreur sur les pas d'itération reste supérieure à (3.27), le critère d'arrêt étant leur égalité. Algorithmiquement, on procède comme suit. On introduit le vecteur résidu ξ^k défini par la formule

$$\xi^k = \Lambda^* (\Lambda\varphi^k - f) = \Lambda^*\Lambda (\varphi^k - \varphi). \tag{3.28}$$

On a la majoration

$$\| \xi^k \| \leqslant \| \Lambda^*\Lambda \| \, \| \varphi^k - \varphi \|. \tag{3.29}$$

La quantité $\| \varphi - \varphi_\varepsilon \|$ doit évidemment être équivalente à δ, ce qui conduit à la condition

$$\| \xi^k \| \leqslant \delta \, \| \Lambda^*\Lambda \|, \tag{3.30}$$

i.e. on arrête les calculs quand la norme du résidu $\| \xi^k \|$ devient comparable au second membre de (3.30). Ainsi, on a l'estimation paramétrique pour k_0

$$\| \xi^{k_0} \| \leqslant \beta \, (\Lambda^*\Lambda) \, \delta. \tag{3.31}$$

Ainsi, on n'aborde les problèmes d'évolution inverses qu'après avoir étudié avec soin divers modèles simples qui fournissent l'information à priori nécessaire pour construire un algorithme numérique de qualité. Des cas plus ardus peuvent se présenter, mais l'analyse faite suffit pour se faire une idée de certains principes dont on part pour construire les méthodes numériques compte tenu des erreurs survenues et des modèles simples. Nous avons discuté une seule approche de la régularisation, qui renseigne cependant assez

bien sur les tactiques numériques employées dans les problèmes inverses. Pour une étude plus poussée, nous renvoyons à la monographie de M.M. Lavrentiev [II].

Nous tenons enfin à dire que les méthodes et les idées exposées jouent pour le problème de Cauchy relatif aux équations elliptiques. Il s'agit des problèmes mal posés au sens classique, si bien qu'on s'adresse à des méthodes élaborées en vue des problèmes conditionnellement bien posés. Parmi les auteurs qui se sont distingués dans ce domaine, citons Tikhonov [XVI ; 17], M.M. Lavrentiev [XVI ; 5, 25] et Ivanov [XVI ; 1].

7.4. Problèmes inverses posés en termes de la théorie des perturbations

On recourt de plus en plus volontiers, pour former les algorithmes de calcul, à des problèmes inverses posés en termes de fonctions adjointes et de méthodes de la théorie des perturbations. C'est surtout vrai des problèmes complexes de la physique mathématique dans lesquels on apprécie difficilement à priori diverses influences exercées sur la solution. La chose s'avère particulièrement importante dans la planification des expériences lorsqu'on se propose d'obtenir un ensemble de fonctionnelles qui soit le plus riche du point de vue informationnel.

Fuks [XVI], Oussatchev [XVI ; 18], Kadomtsev [XVI ; 2], Marchouk, Orlov [XVI ; 14] ont obtenu des résultats de valeur relatifs aux problèmes inverses en théorie du transfert de rayonnement. La position et la résolution des problèmes inverses interviennent de façon sensible dans les problèmes de reconnaissance des formes, dans l'identification et la théorie de l'optimisation. Ces questions sont étudiées à fond par Pontriaguine et al. [III], Balakrishnan, Neustadt [III], Lions [II].

7.4.1. Certaines questions de la théorie linéaire des mesures

La théorie des mesures s'insère aujourd'hui dans la gestion du fonctionnement des systèmes informatiques. Les appareils de mesure fournissent des renseignements (des fonctionnelles) sur le phénomène, ce qui permet d'analyser celui-ci et de l'orienter dans un sens voulu. Ces fonctionnelles aidant, on dégage l'interprétation du phénomène physique.

Nous ne parlerons pas de mesures isolées telles que la mesure de tension ou la mesure d'intensité du courant en tel point du circuit électrique. Seuls nous intéressent les phénomènes et les transformations physiques complexes qui doivent être compris et appréciés quantitativement avec une précision désirée. Des problèmes analogues se posent fréquemment, surtout dans les secteurs nouveaux-nés de la technique. On ne saurait par exemple élaborer des méthodes pour mesurer le facteur de multiplication du réacteur si l'on ignore

un détail quelconque de la réaction en chaîne, de la diffusion de neutrons ou des équations qui gouvernent le fonctionnement du réacteur dans des conditions différentes.

Il va sans dire que les méthodes de mesure et les appareils mêmes se perfectionnent en même temps que la théorie du phénomène physique impliqué. La mise au point de la théorie et des expériences va en général de pair avec l'amélioration des procédés existants et l'apparition de méthodes nouvelles.

Une question se pose: ne peut-on formuler une approche plus ou moins générale des méthodes de mesure pour des processus différents qui permette une description mathématique formelle de l'algorithme. La chose est en effet possible au moins en ce qui concerne les problèmes à opérateur linéaire. C'est cette classe de problèmes qui sera l'objet de notre étude.

Il nous paraît que la théorie des variations mesurées des grandeurs physiques peut reposer sur la théorie des perturbations. Supposons que nous étudions un phénomène physique complexe à l'aide d'un instrument à caractéristiques physiques déterminées. Ses enregistrements liés au champ étudié de la grandeur physique concernée sont des fonctionnelles de ce champ. Dans la plupart des cas, l'expérimentateur néglige cependant celui-ci car il veut connaître les écarts par rapport au champ induits par les perturbations qui sont en général petites. Les mesures doivent donc être effectuées avec une précision suffisante pour qu'on enregistre les écarts à un état « standard ». Supposons qu'il en est ainsi. L'information ainsi obtenue suffit-elle pour bien interpréter l'expérience et peut-on reconstituer de façon assez exacte l'information sur l'état perturbé du système? On n'y répond en général pas aisément car la reconstitution de l'information sur le champ d'une grandeur physique par les instruments de mesure constitue de règle un problème mal posé.

On lève cette difficulté fondamentale en établissant dès le début une relation entre les déviations de l'appareil et les écarts des paramètres physiques étudiés. Dans ce cas, l'erreur sur la caractéristique concernée est proportionnelle à l'erreur dans les enregistrements (la variation de fonctionnelle), si bien qu'on interprète l'expérience moyennant le plus d'information possible. C'est avec ces concepts que nous allons traiter le côté théorique de la question. (Nous nous appuierons essentiellement sur les résultats de Marchouk et Orlov [XVI; 14].)

7.4.2. Fonctions adjointes et notion de fonction d'influence

Considérons une fonction $\varphi(x)$ telle que

$$L\varphi(x) = q(x), \tag{4.1}$$

L étant un opérateur linéaire et $q(x)$ la répartition de sources dans le milieu. On définit x comme étant l'ensemble des variables du

problème (coordonnées de temps et d'espace, énergie, direction de la vitesse) et on estime que φ et q sont réels et $\varphi \in \Phi$.

Supposons, pour fixer les idées, que le phénomène étudié a trait à la diffusion ou au transport de substance bien que les résultats théoriques dépassent le cadre de ces problèmes.

On introduit l'espace de Hilbert des fonctions à produit scalaire

$$(g, h) = \int g(x) h(x) dx, \tag{4.2}$$

l'intégration étant étendue à tout le domaine de définition D des fonctions g et h.

Dans les problèmes physiques on cherche en général la valeur d'une quantité qui est une fonctionnelle de $\varphi(x)$. Toute quantité liée linéairement à $\varphi(x)$ peut être représentée par ce produit scalaire. Si l'on demande par exemple la mesure d'un phénomène par un appareil à caractéristique $\Sigma(x)$, cette valeur est

$$J_\Sigma[\varphi] = \int \varphi(x) \Sigma(x) dx = (\varphi, \Sigma). \tag{4.3}$$

On considérera donc les quantités qui s'expriment par des fonctionnelles linéaires de $\varphi(x)$:

$$J_p[\varphi] = (\varphi, p),$$

où p caractérise le phénomène physique concerné. Introduisons l'adjoint de L par la formule

$$(g, Lh) = (h, L^*g) \tag{4.4}$$

quels que soient g et h. On associe à (4.1) l'équation adjointe formellement non homogène

$$L^*\varphi_p^* = p(x), \tag{4.5}$$

avec $p(x)$ une fonction arbitraire et $\varphi_p^* \in \Phi^*$. Substituons à h et à g de (4.4) les solutions φ et φ_p^* des équations (4.1) et (4.5), il vient

$$(\varphi_p^*, L\varphi) = (\varphi, L^*\varphi_p^*) \tag{4.6}$$

ou, compte tenu de (4.1) et (4.5),

$$(\varphi_p^*, q) = (\varphi, p); \tag{4.7}$$

autrement dit, $J_q[\varphi_p^*] = J_p[\varphi]$. On cherche donc la valeur de la fonctionnelle $J_p[\varphi]$ ou bien en résolvant (4.1) et en utilisant

$$J_p[\varphi] = (\varphi, p), \tag{4.8}$$

ou bien en résolvant (4.5) et en définissant $J_p[\varphi]$ par la formule

$$J_p[\varphi] = J_q[\varphi_p^*] = (\varphi_p^*, q). \tag{4.9}$$

On associe donc à chaque fonctionnelle linéaire $J_p[\varphi] = (\varphi, p)$ une fonction $\varphi_p^*(x)$ vérifiant (4.5) dont le second membre doit être la fonction $p(x)$ caractéristique du phénomène étudié.

Soit, dans le milieu, une « source de puissance unité » x_0, i.e.

$$q(x) = \delta(x - x_0). \tag{4.10}$$

Comme

$$(\varphi(x), \delta(x - x_0)) = \varphi(x_0), \tag{4.11}$$

on a

$$J_p[\varphi] = J_{q=\delta(x-x_0)}[\varphi_p^*] = \varphi_p^*(x_0). \tag{4.12}$$

Par conséquent, la fonction adjointe $\varphi_p^*(x)$ décrit la relation entre la fonctionnelle $J_p[\varphi] = (\varphi, p)$ et le point x_0.

Représentons en pensée un système physique (ou un appareil) qui mesure une quantité $J_p[\varphi]$, fonctionnelle linéaire de la solution, qui dépend par exemple de la densité de particules φ. Si un point du système reçoit une certaine quantité de particules (ou s'il en perd), la valeur mesurée de $J_p[\varphi]$ croît (ou décroît) et cette variation dépend du point concerné. On sait que cette dépendance obéit à la fonction adjointe $\varphi_p^*(x)$ satisfaisant à (4.5). On peut donc dire que $\varphi_p^*(x)$ représente la contribution d'une particule à la fonctionnelle J_p, d'où son nom de fonction d'influence (ou fonction d'importance).

Avec la fonction adjointe $\varphi_p^*(x)$ ainsi définie, la théorie des perturbations pour toute fonctionnelle $J_p[\varphi]$ se prête à une interprétation simple. En effet, si l'élément de volume Δx centré en x reçoit ou perd δN particules, la variation correspondante de J_p s'exprime par

$$\delta J_p = \delta N \varphi_p^*(x). \tag{4.13}$$

Si l'on modifie légèrement les paramètres de notre système de façon que l'opérateur L devient $L + \delta L$, la quantité de particules dans chaque élément de volume Δx varie de $\delta N = -\Delta x \delta L \varphi$ et la variation totale de J_p s'écrit

$$\delta J_p = -\int \varphi_p^*(x)\, \delta L \varphi(x)\, dx. \tag{4.14}$$

La démonstration rigoureuse de cette égalité sera faite plus loin.

La relation (4.13) permet de mesurer la répartition de la fonction d'influence dans le système en modifiant d'une façon déterminée le nombre de particules en des points x différents et en mesurant la variation correspondante de J_p. La notion de fonction d'influence peut être utile en théorie de divers instruments de mesure. En effet, il y a en général autant d'appareils de mesure que de quantités J_p et on introduit à chaque fois $\varphi_p^*(x)$ bien définie qu'on calcule ou mesure une seule fois. Connaissant la répartition de la substance et de sa fonction d'influence, on fait de la relation (4.14) deux coups.

Primo, en mesurant δJ_p pour δL différents, on détermine par (4.14) ces dernières quantités, i.e. les caractéristiques de l'interaction des particules avec la substance. On peut par exemple mesurer (et on le fait effectivement) les sections pour l'interaction des neutrons avec la matière pour divers échantillons : on met les échantillons dans l'appareil de mesure et on définit $\delta\Sigma = \delta L$ en fonction de la variation de J_p.

Secundo, la relation (4.14) permet d'effectuer des corrections sur J_p grâce à des éléments perturbateurs de l'instrument de mesure.

On note enfin que la notion de fonction d'influence permet d'établir les équations pour $\varphi_p^*(x)$ à la lumière de la signification physique de cette quantité tout comme l'équation du flux neutronique s'obtient de la loi de conservation du nombre de neutrons.

Les formules ci-dessus conduisent de plus à un théorème de réciprocité pour les fonctions de Green $G(x, x_0)$, $G^*(x, x_1)$. La première fonction vérifie l'équation (4.1) pour $q(x) = \delta(x - x_0)$ et la deuxième satisfait à (4.5) si $p(x) = \delta(x - x_1)$.

Portons dans (4.7) $\varphi(x) = G(x, x_0)$, $\varphi_p^*(x) = G^*(x, x_1)$ et ces expressions de q et p, il vient

$$G(x_1, x_0) = G^*(x_0, x_1), \tag{4.15}$$

ce qui exprime le théorème de réciprocité.

7.4.3. Théorie des perturbations pour les fonctionnelles linéaires

Si le milieu en interaction avec le champ varie, i.e. si l'opérateur de l'équation (4.1) devient

$$L' = L + \delta L,$$

le champ de $\varphi(x)$ et la fonctionnelle $J_p[\varphi]$ changent eux aussi :

$$\varphi(x) \to \varphi'(x), \quad J_p[\varphi] \to J_p' = J_p + \delta J_p.$$

On établit la relation entre δL et δJ_p.

Le système perturbé obéit à l'équation

$$L'\varphi' = (L + \delta L)\varphi' = q \tag{4.16}$$

et la fonction adjointe associée à J_p du système non perturbé à

$$L^*\varphi_p^* = p. \tag{4.17}$$

On multiplie scalairement (4.16) par φ^*, l'équation (4.17) par φ' et on fait la soustraction, il vient à gauche par définition de l'opérateur adjoint (4.4)

$$(\varphi_p^*, L'\varphi') - (\varphi', L^*\varphi_p^*) = (\varphi_p^*, \delta L\varphi'), \tag{4.18}$$

et à droite, conformément à (4.7),

$$(\varphi_p^*, q) - (\varphi', p) = J_p[\varphi] - J_p[\varphi'] = -\delta J_p. \tag{4.19}$$

En égalant les deux expressions, on obtient la formule générale pour l'accroissement de fonctionnelle

$$\delta J_p = -(\varphi_p^*, \delta L\varphi'). \tag{4.20}$$

Si l'on prend, au lieu de (4.16) et (4.17), respectivement l'équation adjointe perturbée

$$(L^* + \delta L^*)\,\varphi_p^{*\prime} = p \tag{4.21}$$

et l'équation non perturbée (4.1), on trouve de même la relation

$$\delta J_p = -(\varphi, \delta L^*\varphi_p^{*\prime}) \tag{4.22}$$

qui est évidemment équivalente à (4.20).

Chose importante, les formules de la théorie des perturbations s'appliquent ici à la variation de fonctionnelle dont l'erreur est d'ordinaire tolérable à plusieurs pour cent près, si bien qu'on calcule ces variations sans connaître la solution exacte du problème proposé et de son adjoint, les solutions approchées faisant parfaitement l'affaire.

Si la perturbation affectant l'opérateur L (resp. L^*) est suffisamment faible pour que φ (resp. φ_p^*) soit peu altérée, on effectue dans (4.20) (resp. (4.22)) la substitution approchée $\varphi' \approx \varphi$ (resp. $\varphi^{*\prime} \approx \varphi^*$) et on obtient deux formules équivalentes de la théorie des petites perturbations *)

$$\delta J_p = -(\varphi_p^*, \delta L\varphi), \tag{4.23}$$

$$\delta J_p = -(\varphi, \delta L^*\varphi_p^*). \tag{4.24}$$

A part qu'elles permettent d'apprécier divers effets et d'analyser les mesures, ces formules trouvent une autre application fort importante.

En théorie et dans les problèmes pratiques, on remplace souvent le système complexe considéré par un modèle simplifié. Une condition nécessaire en est évidemment que cette transformation conserve certaines caractéristiques essentielles du système. On remplace par exemple les coefficients variables des équations différentielles par des constantes ou encore les conditions aux limites réelles par d'autres plus simples qui donnent la valeur vraie d'une fonctionnelle choisie (méthode des conditions aux limites efficaces).

Nous avons obtenu les formules de la théorie des perturbations qui constituent une approche très générale de divers problèmes. Supposons que le système considéré est caractérisé par l'opérateur L

*) Pour simplifier l'exposé, nous dirons dans la suite que (4.20) et (4.22) sont les formules de la théorie des perturbations.

et que la caractéristique essentielle est la fonctionnelle $J_p[\varphi]$. Si le modèle simple cherché a pour opérateur $L' = L + \delta L$, alors J_p est conservé par le passage au modèle sous la condition nécessaire

$$\delta J_p = -(\varphi_p^*, [L' - L]\varphi') = 0, \text{ i.e. } (\varphi_p^*, L'\varphi') = (\varphi_p^*, L\varphi'). \quad (4.25)$$

Si l'on s'intéresse à plusieurs quantités $J_{p_1}, J_{p_2}, \ldots$, on a autant de conditions (4.25) avec les solutions $\varphi_{p_1}^*, \varphi_{p_2}^*, \ldots$

La condition (4.25) ne définit pas univoquement le modèle équivalent. Ce n'est qu'une condition nécessaire de sa définition. Si l'opérateur L' déterminé par des considérations physiques contient un ou plusieurs paramètres, on utilise en particulier (4.25) pour en définir les valeurs.

7.4.4. Méthodes numériques de résolution des problèmes inverses et planification des expériences

Etant donné un ensemble de fonctionnelles (mesures) J_{p_i}, $i = 1, 2, \ldots, n$, on admet que les mesures sont de nature variée: on effectue les mesures en des « points » différents du domaine de définition de la solution ou on enregistre des caractéristiques différentes du phénomène étudié. On suppose, pour simplifier l'exposé, que les erreurs de mesures statistiques sont corrigées, i.e. on a effectué un traitement partiel des données.

On associe à chaque fonctionnelle J_{p_i} une fonction d'influence pour le problème non perturbé (le modèle) où l'opérateur L et son domaine sont supposés connus. On a à résoudre n problèmes

$$L^*\varphi_{p_i} = p_i, \qquad i = 1, 2, \ldots, n. \quad (4.26)$$

On cherche au préalable n fonctions $\varphi_{p_i}^*$ et on aborde le problème à L « non perturbé » adjoint de L^*

$$L\varphi = q. \quad (4.27)$$

On fait l'hypothèse « $\varphi \in \Phi$ et $\varphi^* \in \Phi^*$ », Φ et Φ^* étant les domaines respectifs de L et L^*. On construit n formules de la théorie des petites perturbations

$$(\varphi_{p_i}^*, \delta L\varphi) = -\delta J_{p_i}, \qquad i = 1, 2, \ldots, n, \quad (4.28)$$

avec δL la différence entre l'opérateur étudié L' et l'opérateur modèle L. L'opérateur L est supposé connu:

$$L = \sum_{k=1}^{m} [\alpha_k A_k + B_k(\beta_k C_k)], \quad (4.29)$$

où A_k, B_k et C_k sont des opérateurs linéaires élémentaires, par exemple les opérateurs de dérivation ou d'intégration, ou encore

leurs combinaisons; $\alpha_k(x)$ et $\beta_k(x)$ les coefficients à déterminer dont on connaît d'ordinaire une approximation grossière pour le problème non perturbé (modèle).

On se propose maintenant de reconstituer α'_k et β'_k dans

$$L' = \sum_{k=1}^{m} [\alpha'_k A_k + B_k(\beta'_k C_k)]. \tag{4.30}$$

On trouve moyennant (4.29), (4.30)

$$\delta L = \sum_{k=1}^{m} [\delta\alpha_k A_k + B_k(\delta\beta_k C_k)], \tag{4.31}$$

où

$$\delta\alpha_k = \alpha'_k - \alpha_k, \quad \delta\beta_k = \beta'_k - \beta_k.$$

Le report de (4.31) dans (4.28) donne sous des conditions appropriées le système d'équations

$$\sum_{k=1}^{m} [(\varphi^*_{p_i}, \delta\alpha_k A_k\varphi) + (B^*_k\varphi^*_{p_i}, \delta\beta_k C_k\varphi)] = -\delta J_{p_i}, \tag{4.32}$$

$$i = 1, 2, \ldots, n.$$

Le problème suivant est de paramétriser les variations $\delta\alpha_k$ et $\delta\beta_k$. Plaçons-nous dans le cas simple où $\delta\beta_k = 0$ et $\delta\alpha_k$ sont des constantes. Le système (4.32) devient alors le problème d'algèbre linéaire

$$\sum_{k=1}^{m} \delta\alpha_k(\varphi^*_{p_i}, A_k\varphi) = -\delta J_{p_i}, \qquad i = 1, 2, \ldots, n. \tag{4.33}$$

Ici $(\varphi^*_{p_i}, A_k\varphi)$ sont les éléments matriciels qu'on calcule pour φ, $\varphi^*_{p_i}$ et A_k donnés.

Soit y un vecteur de composantes $\delta\alpha_k$, F un vecteur formé de δJ_{p_i} et $a_{ik} = (\varphi^*_{p_i}, A_k\varphi)$ les éléments d'une matrice Λ. On a l'équation

$$\Lambda y = F. \tag{4.34}$$

Si le nombre n de fonctionnelles coïncide avec celui des variations cherchées des coefficients α_k, alors $\delta\alpha_k$ s'obtient en principe à l'aide de ce système. Si $n > m$, le système (4.34) se trouve surdéterminé et sa solution (quand elle existe) est d'habitude cherchée par la méthode des moindres carrés avec la condition que y réalise le minimum de la fonctionnelle quadratique

$$\| \Lambda y - F \|^2 = \min. \tag{4.35}$$

On dit des fois que le vecteur y est quasi-solution de (4.34). Lorsque $n = m$, on aborde (4.34) par des procédés étudiés plus haut (Chap. 4)

en rapport avec les processus itératifs avec des données initiales inexactes.

Si $\delta\alpha_k(x)$ et $\delta\beta_k(x)$ sont des fonctions, on cherche la solution du problème inverse par des méthodes de paramétrisation. Voici l'idée de ces méthodes. On suppose qu'une analyse à priori (calculs statistiques, corrélation) des paramètres physiques conduit à des systèmes orthogonaux complets de fonctions $u_{k,l}(x)$ et $v_{k,l}(x)$ tels qu'on approche assez bien, pour $n(k)$ petit, les fonctions α_k et β_k, de sorte que

$$\begin{aligned} \delta\alpha_k(x) &= \sum_{l=1}^{n(k)} a_{k,l}\, u_{k,l}(x), \\ \delta\beta_k(x) &= \sum_{l=1}^{n(k)} b_{k,l}\, v_{k,l}(x), \end{aligned} \tag{4.36}$$

$a_{k,l}$ et $b_{k,l}$ étant les coefficients à déterminer.

La substitution de (4.36) dans (4.32) donne

$$\sum_{k=1}^{m}\sum_{l=1}^{n(k)} [a_{k,l}(\varphi^*_{p_i},\ u_{k,l}A_k\varphi) + b_{k,l}(B^*_k\varphi^*_{p_i},\ v_{k,l}C_k\varphi)] = -\delta J_{p_i}, \quad i=1, 2, \ldots, n, \tag{4.37}$$

Ordonnons $a_{k,l}, b_{k,l}$ et notons-les y_j, $j = 1, 2, \ldots$ On introduit une matrice Λ telle que l'équation

$$\Lambda y = F$$

soit équivalente au système (4.37). On est conduit une fois de plus au problème d'algèbre linéaire (4.37) qu'on résout par rapport à $a_{k,l}$, $b_{k,l}$, ce qui donne respectivement $\delta\alpha_k$ et $\delta\beta_k$.

Nous nous sommes bornés au cas où la solution du problème modèle approche bien la solution réelle, i.e. on remplace φ' par φ en faisant appel à la théorie des petites perturbations. Si l'état non perturbé (modèle) du processus diffère de façon sensible de l'état réel, l'algorithme considéré constitue une première approximation de la solution du problème inverse. Une fois les variations $\delta\alpha_k$ et $\delta\beta_k$ obtenues, on corrige les coefficients α_k, β_k et on calcule

$$\begin{aligned} \alpha'_k &= \alpha_k + \delta\alpha_k, \\ \beta'_k &= \beta_k + \delta\beta_k. \end{aligned}$$

On passe ensuite au problème « perturbé »

$$L'\varphi' = f, \tag{4.38}$$

où l'opérateur L' se définit par

$$L' = \sum_{k=1}^{m} [\alpha'_k A_k + B_k(\beta'_k C_k)],$$

on réalise l'approximation suivante de la solution en prenant, au lieu de (4.32) la formule plus générale des perturbations

$$\sum_{k=1}^{m} [(\varphi^*_{p_i},\ \delta\alpha_k A_k \varphi') + (B^*_k \varphi^*_{p_i},\ \delta\beta_k C_k \varphi')] = -\delta J_{p_i}, \qquad (4.39)$$
$$i = 1,\ 2,\ \ldots,\ n,$$

et on recommence l'algorithme pour préciser $\delta\alpha_k$, $\delta\beta_k$ (on réalise ainsi une seconde approximation de la solution du problème inverse). Le processus peut évidemment continuer. On démontre la convergence des approximations successives à la lumière d'une information concrète sur les opérateurs élémentaires A_k et les domaines de L et L^*.

Illustrons l'algorithme décrit sur un exemple simple. Soit le problème

$$-\frac{d}{dx}\beta(x)\frac{d\varphi'}{dx} + \alpha(x)\varphi' = f(x), \qquad \varphi'(0) = \varphi'(1) = 0 \qquad (4.40)$$

à coefficients $\alpha(x)$, $\beta(x)$ inconnus qu'on suppose à priori être des fonctions continues dans le domaine de définition $0 \leqslant x \leqslant 1$ de la solution et dont on connaît des approximations $\bar{\alpha}$ et $\bar{\beta}$, i.e.

$$\alpha(x) = \bar{\alpha} + \delta\alpha(x), \qquad \beta(x) = \bar{\beta} + \delta\beta(x). \qquad (4.41)$$

Si l'information à priori permet de choisir $\alpha(x)$, $\beta(x)$ du modèle plus précis, l'hypothèse de leur égalité à des constantes $\bar{\alpha}$ et $\bar{\beta}$ n'est nullement obligatoire.

Une étude préliminaire fait conclure de plus à la possibilité de représenter $\delta\alpha(x)$ et $\delta\beta(x)$ par les sommes finies

$$\delta\alpha(x) = \sum_{l=1}^{n(1)} a_l u_l(x), \qquad \delta\beta(x) = \sum_{l=1}^{n(1)} b_l v_l(x), \qquad (4.42)$$

avec $\{u_l(x)\}$ et $\{v_l(x)\}$ des systèmes orthonormés complets de fonctions (cela peut être par exemple les polynômes de Legendre, les polynômes trigonométriques).

Soit $p_1(x)$, $p_2(x)$, ..., $p_n(x)$ des caractéristiques des mesures telles qu'à chaque fois on enregistre la fonctionnelle

$$J'_{p_i}[\varphi'] = \int_0^1 p_i(x)\varphi'(x)\,dx, \qquad i = 1,\ 2,\ \ldots,\ n. \qquad (4.43)$$

On dit des fonctions $p_i(x)$ qu'elles sont les caractéristiques de l'appareil pour une mesure donnée.

On introduit le problème non perturbé (modèle) associé à (4.40)

$$-\frac{d}{dx}\bar{\beta}\frac{d\varphi}{dx} + \bar{\alpha}\varphi = f, \qquad \varphi(0) = \varphi(1) = 0, \qquad (4.44)$$

et on formule n problèmes adjoints pour le modèle choisi

$$-\frac{d}{dx}\bar{\beta}\frac{d\varphi^*_{p_i}}{dx}+\bar{\alpha}\varphi^*_{p_i}=p_i(x), \qquad \varphi^*_{p_i}(0)=\varphi^*_{p_i}(1)=0, \quad i=1, 2, \ldots, n. \tag{4.45}$$

La théorie générale veut que

$$J_{p_i}[\varphi]=\int_0^1 p_i(x)\,\varphi(x)\,dx=\int_0^1 f(x)\,\varphi^*_{p_i}(x)\,dx. \tag{4.46}$$

On suppose résolus les problèmes modèles (4.44) et (4.45) et on cherche les variations δJ_{p_i} par la formule

$$\delta J_{p_i}=J'_{p_i}-J_{p_i}, \quad i=1, 2, \ldots, n, \tag{4.47}$$

où J'_{p_i} est l'enregistrement de l'appareil caractérisé par p_i, qui correspond à (4.43) (avec φ' inconnu) et J_{p_i} une fonctionnelle qu'on recherche théoriquement par toute relation (4.46). Les mesures doivent être effectuées avec une précision suffisante pour qu'on calcule δJ_{p_i}.

Passons aux formules de la théorie des petites perturbations (4.33) :

$$A=E, \quad B=-\frac{d}{dx}, \quad C=\frac{d}{dx}.$$

On a, compte tenu des conditions aux limites pour $\varphi^*_{p_i}$ et φ,

$$\int_0^1\left(\delta\alpha\varphi\varphi^*_{p_i}+\delta\beta\frac{d\varphi}{dx}\frac{d\varphi^*_{p_i}}{dx}\right)dx=-\delta J_{p_i}. \tag{4.48}$$

Portons dans cette formule les valeurs (4.42) de $\delta\alpha\,(x)$ et $\delta\beta\,(x)$, il vient

$$\sum_{l=1}^{n(1)}\left(a_l\int_0^1 u_l\varphi\varphi^*_{p_i}\,dx+b_l\int_0^1 v_l\frac{d\varphi}{dx}\frac{d\varphi^*_{p_i}}{dx}dx\right)=-\delta J_{p_i}, \tag{4.49}$$

$$i=1, \quad 2, \ldots, n.$$

Le système se trouve complètement défini si $n=2n$ (1).

On résout ce système par rapport à a_l, b_l et on obtient, moyennant la représentation (4.42), une première approximation de α' et β'. On améliore les résultats par la technique ci-dessus des approximations successives. On procède de même en ce qui concerne les problèmes inverses plus complexes dont le cas des perturbations δf affectant les sources.

Ainsi, on est conduit au problème de planifier une expérience complexe qui consiste à choisir, parmi toutes les mesures possibles

(réalisables pratiquement), celles qui donnent la plus utile information du point de vue de la résolution du problème inverse : reconstituer certaines caractéristiques du milieu (les coefficients des équations). S'agissant de l'optimisation globale, ce problème s'avère fort délicat. Aussi nous nous bornerons à certaines approches particulières.

Faisons l'hypothèse suivante : avant l'expérience, on construit un modèle du problème non perturbé, on décrit moyennant ce modèle les fonctionnelles linéaires de la solution et on utilise l'information à priori sur la précision des mesures pour établir celle à imposer aux mesures des fonctionnelles. Supposons garanties les conditions nécessaires pour que les fonctionnelles δJ_{p_i} soient mesurées avec la précision voulue. Parmi les ensembles d'enregistrements on choisit ceux qui conduisent à la matrice Λ la mieux conditionnée. Le système linéaire obtenu est résolu sans difficulté, si bien que de toutes les tactiques données, cette façon de mener l'expérience est optimale dans un sens (il est vrai qu'on laisse de côté les aspects économiques de la planification qui s'avèrent des fois décisifs). Si l'ensemble donné de fonctionnelles porteuses d'information qui garantissent le meilleur conditionnement de Λ, ne permet pas d'atteindre la précision imposée sur J_{p_i} mesurées, un problème plus complexe se pose : planifier l'expérience avec des contraintes données sur la précision des mesures permises par le pouvoir de résolution de l'appareillage. C'est en fait une sorte de problème d'optimisation avec contraintes.

Nous estimons inutile de nous arrêter sur le traitement statistique des données expérimentales, car, *primo*, il fait l'objet de nombreux ouvrages et, *secundo*, il n'est pas de nature à compliquer de façon quelque peu sérieuse la théorie des problèmes inverses (voir Marchouk, Drobychev [XVI ; 13]).

7.5. Théorie des perturbations et modèles non linéaires complexes

La construction d'un modèle mathématique de processus et phénomènes complexes constitue en général une entreprise difficile car ces modèles sont tenus de prendre en considération nombre d'effets variés dont tous ne sont pas décrits avec une précision imposée. On recourt donc à tel instant à telle formulation mathématique simplifiée et on décrit plusieurs caractéristiques du phénomène en négligeant divers aspects parfois très importants. Malgré les défauts mentionnés, les modèles représentent en général le processus physique avec la fidélité voulue. Quant aux effets négligés, on les apprécie à l'aide d'une théorie spéciale des perturbations (voir numéro précédent).

Le présent numéro traitera de façon assez générale la construction et l'analyse des modèles mathématiques. Contrairement à l'exposé ci-dessus, on s'occupera des processus gouvernés par les équations non linéaires. On montrera que certaines approximations de processus non linéaires peuvent être abordées et analysées par des méthodes propres aux problèmes linéaires. Il va sans dire que ces approches s'appuient sur divers procédés de linéarisation.

7.5.1. Equations et leurs adjointes

Soit un processus stationnaire qui obéit à l'équation opératorielle

$$A\varphi = f, \tag{5.1}$$

où A est un opérateur différentiel matriciel dépendant de la solution fonction vectorielle φ et des données initiales $\alpha_1, \alpha_2, \ldots, \alpha_n$, fonctions de coordonnées, et le vecteur donné f représente les sources (fonctions de coordonnées) et dépend des paramètres donnés $\beta_1, \beta_2, \ldots, \beta_m$, i.e. $A = A(\varphi, \alpha_1, \alpha_2, \ldots, \alpha_n)$ et $f = f(\beta_1, \beta_2, \ldots, \beta_m)$.

Soit $\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_n$ et $\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_m$ des données relatives à un état standard du système, qui sera dit non perturbé. L'état non perturbé obéit naturellement à l'équation

$$A(\bar{\varphi}, \bar{\alpha}_1, \ldots, \bar{\alpha}_n)\,\bar{\varphi} = f(\bar{\beta}_1, \ldots, \bar{\beta}_m). \tag{5.2}$$

Avec les notations

$$\bar{A} = A(\bar{\varphi}, \bar{\alpha}_1, \ldots, \bar{\alpha}_n), \qquad \bar{f} = f(\bar{\beta}_1, \ldots, \bar{\beta}_m),$$

l'équation (5.2) se récrit

$$A\varphi = \bar{f}. \tag{5.3}$$

On suppose connue la solution de (5.3).

On admet ensuite que l'état réel (perturbé) du système est gouverné par l'équation (5.1) dont les entrées ne s'écartent que fort peu des données de l'état standard, i.e.

$$\alpha_i = \bar{\alpha}_i + \delta\alpha_i, \qquad \beta_j = \bar{\beta}_j + \delta\beta_j,$$

les écarts $\delta\alpha_i$ et $\delta\beta_j$ étant supposés petits devant $\bar{\alpha}_i$ et $\bar{\beta}_j$ respectivement.

Dans ce cas, on a au lieu de (5.2):

$$A(\bar{\varphi} + \delta\varphi, \bar{\alpha}_i + \delta\alpha_i)(\bar{\varphi} + \delta\varphi) = f(\bar{\beta}_j + \delta\beta_j), \tag{5.4}$$

où

$$\varphi = \bar{\varphi} + \delta\varphi.$$

On suppose de même à priori que $\delta\varphi$ est très petit devant $\bar{\varphi}$. Sous l'hypothèse de A, de la solution φ et des entrées suffisamment régu-

liers, on considère les développements

$$\begin{aligned} A(\bar{\varphi}+\delta\varphi,\ \bar{\alpha}_i+\delta\alpha_i) &= \bar{A}+\frac{\overline{\partial A}}{\partial\varphi}\delta\varphi+\frac{\overline{\partial A}}{\partial\alpha_i}\delta\alpha_i+\ldots, \\ f(\bar{\beta}_j+\delta\beta_j) &= \bar{f}+\frac{\overline{\partial f}}{\partial\beta_j}\delta\beta_j. \end{aligned} \tag{5.5}$$

Portons dans (5.4) et arrêtons-nous aux termes d'ordre 1, il vient

$$\bar{A}\bar{\varphi}+\left(\bar{A}+\frac{\overline{\partial A}}{\partial\varphi}\bar{\varphi}\right)\delta\varphi+\frac{\overline{\partial A}}{\partial\alpha_i}\bar{\varphi}\,\delta\alpha_i=\bar{f}+\frac{\overline{\partial f}}{\partial\beta_j}\delta\beta_j, \tag{5.6}$$

et, compte tenu de (5.3),

$$\left(\bar{A}+\frac{\overline{\partial A}}{\partial\varphi}\bar{\varphi}\right)\delta\varphi=\frac{\overline{\partial f}}{\partial\beta_j}\delta\beta_j-\frac{\overline{\partial A}}{\partial\alpha_i}\bar{\varphi}\delta\alpha_i. \tag{5.7}$$

Cette équation définit les faibles écarts de la solution φ par rapport à l'état non perturbé.

Supposons maintenant que l'opérateur état non perturbé $\bar{A}$ et la source $\bar{f}$ sont connus de façon approchée, à savoir

$$\begin{aligned} \bar{A} &= \bar{\Lambda}+\varepsilon, \\ \bar{f} &= \bar{F}+\xi. \end{aligned} \tag{5.8}$$

Ici ε est l'opérateur erreur du modèle:

$$\varepsilon = \bar{A} - \Lambda$$

et ξ l'erreur sur la fonction vectorielle de sources

$$\xi = \bar{f} - \bar{F}.$$

Soit

$$\begin{aligned} \|\bar{A}\| &\gg \|\varepsilon\|, \\ \|\bar{f}\| &\gg \|\xi\|. \end{aligned} \tag{5.9}$$

Les normes de l'opérateur et de la fonction vectorielle sont définies sur les espaces métriques correspondants.

Le report de (5.8) dans (5.7) donne

$$\left(\bar{\Lambda}+\frac{\overline{\partial\Lambda}}{\partial\varphi}\bar{\varphi}\right)\delta\varphi=\frac{\overline{\partial F}}{\partial\beta_j}\delta\beta_j-\frac{\overline{\partial\Lambda}}{\partial\alpha_i}\bar{\varphi}\,\delta\alpha_i+\eta, \tag{5.10}$$

avec

$$\eta=\frac{\overline{\partial\xi}}{\partial\beta_j}\delta\beta_j-\frac{\overline{\partial\varepsilon}}{\partial\alpha_i}\bar{\varphi}\,\delta\alpha_i-\left(\varepsilon+\frac{\overline{\partial\varepsilon}}{\partial\varphi}\bar{\varphi}\right)\delta\varphi. \tag{5.11}$$

Par conséquent, l'équation (5.10) se récrit sous l'hypothèse (5.9)

$$\left(\overline{\Lambda}+\frac{\overline{\partial\Lambda}}{\partial\varphi}\,\overline{\varphi}\right)\delta\varphi=\frac{\overline{\partial F}}{\partial\beta_j}\,\delta\beta_j-\frac{\overline{\partial\Lambda}}{\partial\alpha_i}\,\overline{\varphi}\,\delta\alpha_i+O\left(\|\varepsilon\|+\|\xi\|\right)$$

et on a donc à des infiniment petits près

$$\left(\overline{\Lambda}+\frac{\overline{\partial\Lambda}}{\partial\varphi}\,\overline{\varphi}\right)\delta\varphi=\frac{\overline{\partial F}}{\partial\beta_j}\,\delta\beta_j-\frac{\overline{\partial\Lambda}}{\partial\alpha_i}\,\overline{\varphi}\,\delta\alpha_i. \tag{5.12}$$

L'avant-dernière équation constitue un modèle qui permet de calculer l'état non perturbé du système lorsque les données initiales varient de $\delta\alpha_i$ et $\delta\beta_j$.

On introduit les notations

$$\begin{aligned} L&=\overline{\Lambda}+\frac{\overline{\partial\Lambda}}{\partial\varphi}\,\overline{\varphi},\\ \delta F&=\frac{\overline{\partial F}}{\partial\beta_j}\,\delta\beta_j-\frac{\overline{\partial\Lambda}}{\partial\alpha_i}\,\overline{\varphi}\,\delta\alpha_i. \end{aligned} \tag{5.13}$$

Finalement,

$$L\,\delta\varphi=\delta F, \tag{5.14}$$

et la solution formelle du problème s'écrit

$$\delta\varphi = L^{-1}\delta F. \tag{5.15}$$

7.5.2. Equations adjointes et théorie des perturbations

La formule (5.15) rend le plus de services quand on définit les écarts de la solution pour un seul ensemble de données initiales. Lorsqu'on se propose d'effectuer des modifications dirigées de l'état du modèle, il y a intérêt à effectuer les calculs tests. En effet, on ne saurait apprécier la sensibilité du modèle à la variation de paramètres ni établir le rapport optimal de ceux-ci sans connaître un grand nombre de solutions différentes. On se propose donc d'élaborer des théories des perturbations plus universelles pour des fonctionnelles données, qui permettent d'examiner le modèle mathématique sous tous ses aspects pour des données initiales différentes.

Comme dans le cas linéaire, on introduit

$$L\,\delta\varphi = \delta F \tag{5.16}$$

et l'équation adjointe

$$L^*\varphi^* = p, \tag{5.17}$$

L^* étant l'adjoint de L au sens de Lagrange:

$$(Lg,\ h) = (g,\ L^*h). \tag{5.18}$$

g et h sont deux éléments de l'espace de Hilbert qui se trouvent respectivement dans le domaine de L et celui de L^*. La fonction p n'est pas définie pour le moment.

Multiplions (5.16) et (5.17) scalairement par φ^* et $\delta\varphi$ respectivement et faisons la différence, il vient

$$(\varphi^*,\ L\,\delta\varphi) - (\delta\varphi,\ L^*\varphi^*) = (\varphi^*,\ \delta F) - (p,\ \delta\varphi). \tag{5.19}$$

Le premier membre est nul par suite de (5.18), si bien que

$$(p,\ \delta\varphi) = (\varphi^*,\ \delta F). \tag{5.20}$$

Soit l'ensemble de fonctionnelles linéaires

$$\delta J_n = (p_n,\ \delta\varphi). \tag{5.21}$$

Si, en particulier, $p_n = \delta\,(x - x_n)$, alors

$$\delta J_n = \delta\varphi\,(x_n) \tag{5.22}$$

dans le cas linéaire aussi. On désignera par φ_n^* la fonction φ^* associée à p_n.

Ainsi, on a grâce à (5.20) l'ensemble de fonctionnelles

$$\delta J_n = (\varphi_n^*,\ \delta F). \tag{5.23}$$

Supposons choisies à l'avance N fonctionnelles $J_1,\ J_2,\ \ldots,\ J_N$ et résolus N problèmes adjoints

$$L^*\varphi_n^* = p_n, \qquad n = 1,\ 2,\ \ldots,\ N. \tag{5.24}$$

Selon (5.23), il n'est nullement besoin de calculer $\delta\varphi$ associés aux jeux divers de paramètres $\delta\alpha_i$ et $\delta\beta_j$ car on calcule directement les valeurs des δJ_n pour toute perturbation sur les données initiales à condition de se donner φ_n^*, $n = 1,\ 2,\ \ldots,\ N$.

7.5.3. Théorie des perturbations et problèmes non stationnaires

Considérons un problème non stationnaire abstrait tel que

$$\frac{\partial\varphi}{\partial t} + A\varphi = f, \qquad \varphi = g \quad \text{pour} \quad t = 0 \tag{5.25}$$

et le problème adjoint

$$-\frac{\partial\varphi^*}{\partial t} + A^*\varphi^* = f^*, \qquad \varphi^* = g^* \quad \text{pour } t = T. \tag{5.26}$$

Ici f^* et g^* sont des fonctions vectorielles non définies qu'on choisira plus tard. On multiplie (5.25), (5.26) respectivement par φ^* et φ, on fait la différence, puis on intègre en t sur l'intervalle $0 \leqslant t \leqslant T$. On a

$$\int_0^T \frac{\partial}{\partial t}(\varphi^*,\ \varphi)\,dt + \int_0^T dt\,[(\varphi^*,\ A\varphi) - (\varphi,\ A^*\varphi^*)] =$$

$$= \int_0^T dt\,[(f,\ \varphi^*) - (f^*,\ \varphi)]. \tag{5.27}$$

Comme

$$(\varphi^*, A\varphi) - (\varphi, A^*\varphi^*) = 0,$$

l'expression (5.27) se récrit compte tenu des conditions initiales correspondantes

$$(g^*, \varphi_T) - (g, \varphi_0^*) = \int_0^T dt\,[(f, \varphi^*) - (f^*, \varphi)], \tag{5.28}$$

avec

$$\varphi_T = \varphi|_{t=T}, \qquad \varphi_0^* = \varphi^*|_{t=0}.$$

On se propose maintenant de chercher une fonctionnelle linéaire de la solution de (5.25) telle que

$$J = (g^*, \varphi_T) + \int_0^T (f^*, \varphi)\,dt. \tag{5.29}$$

L'identité (5.28) aidant, la fonctionnelle s'écrit également

$$J = (g, \varphi_0^*) + \int_0^T (f, \varphi^*)\,dt. \tag{5.30}$$

On suppose perturbées les données initiales de (5.25), i.e. on a $g' = g + \delta g$ et $f' = f + \delta f$, auquel cas l'équation (5.30) fournit la formule pour la variation

$$\delta J = (\delta g, \varphi_0^*) + \int_0^T (\delta f, \varphi^*)\,dt. \tag{5.31}$$

Ainsi, on cherche les écarts de J associés aux variations de données initiales sans qu'on ait à résoudre de nombreux problèmes de la forme

$$\frac{\partial \varphi'}{\partial t} + A\varphi' = f', \qquad \varphi' = g' \quad \text{pour} \quad t = 0, \tag{5.32}$$

avec f' et g' différents. Il suffit d'aborder une seule équation adjointe (5.26) et d'utiliser (5.31). Moyennant la formule (5.31) on construit le problème inverse « trouver δg et δf à l'aide de l'ensemble de fonctionnelles δJ ».

Dans de nombreux cas non stationnaires, on s'intéresse non pas aux fonctionnelles de la solution, mais aux écarts qu'elles présentent. On se ramène alors par la méthode du point 5.5.2 aux équations

$$\frac{\partial}{\partial t}\delta\varphi + L\delta\varphi = \delta F, \qquad \delta\varphi = \delta\varphi_0 \quad \text{pour} \quad t = 0 \tag{5.33}$$

et à leurs adjointes

$$-\frac{\partial \varphi^*}{\partial t} + L^*\varphi^* = p, \quad \varphi^* = \varphi_T^* \quad \text{pour} \quad t = T. \tag{5.34}$$

On procède ensuite comme pour les problèmes inverses non stationnaires.

7.5.4. Méthode spectrale et théorie des perturbations

Lorsqu'on veut définir non pas δJ, mais l'écart par rapport à la solution du problème non perturbé, on traite la théorie des perturbations avec d'autres concepts.

On utilise l'équation (5.14)

$$L\,\delta\varphi = \delta F \tag{5.35}$$

et le problème spectral

$$Lw = \lambda w. \tag{5.36}$$

L'opérateur L n'étant pas auto-adjoint, on considère le problème adjoint de (5.36)

$$L^*w^* = \lambda w^*. \tag{5.37}$$

On suppose que (5.36) définit un système complet de fonctions vectorielles propres $\{w_n\}$ relatives aux valeurs propres $\{\lambda_n\}$ et que le problème (5.37) définit $\{w_n^*\}$ et $\{\lambda_n\}$. Les fonctions propres vérifient alors la condition d'orthogonalité

$$(w_n,\ w_m^*) = 0 \quad \text{pour} \quad m \neq n. \tag{5.38}$$

On effectue la normalisation

$$(w_n,\ w_m^*) = \begin{cases} 1, & n = m, \\ 0, & n \neq m. \end{cases} \tag{5.39}$$

On représente la solution de (5.35) par la série de Fourier suivant les fonctions propres de (5.36) :

$$\delta\varphi = \sum_n \delta\varphi_n w_n. \tag{5.40}$$

De même

$$\delta F = \sum_n \delta F_n w_n, \tag{5.41}$$

où

$$\delta\varphi_n = (\delta\varphi,\ w_n^*), \qquad \delta F_n = (\delta F,\ w_n^*).$$

Portons (5.40), (5.41) dans (5.35) et multiplions scalairement par w_m^* :

$$\lambda_m \delta\varphi_m = \delta F_m, \qquad m = 1,\ 2,\ \ldots \tag{5.42}$$

Ainsi, le problème (5.35) admet pour solution

$$\delta\varphi = \sum_n \frac{\delta F_n}{\lambda_n} w_n, \tag{5.43}$$

ou finalement

$$\delta\varphi = \sum_n \frac{(\delta F,\ w_n^*)}{\lambda_n}\, w_n. \tag{5.44}$$

Ainsi, l'écart de la solution vis-à-vis de l'état standard est représenté par la série de Fourier. Si l'on s'intéresse à la fonctionnelle linéaire même, alors

$$\delta J = (p,\ \delta\varphi), \tag{5.45}$$

auquel cas

$$\delta J = \sum_n \frac{(\delta F,\ w_n^*)}{\lambda_n}\,(p,\ w_n) = (\delta F,\ \varphi^*), \tag{5.46}$$

avec

$$\varphi^* = \sum \frac{(p,\ w_n)}{\lambda_n}\, w_n^*,$$

φ^* étant la solution de (5.17).

On traite de même un problème non stationnaire et on obtient les formules correspondantes de la théorie des perturbations.

CHAPITRE 8

DES PROBLÈMES DE LA PHYSIQUE MATHÉMATIQUE

Dans ce chapitre, nous aborderons des problèmes simples types de la physique mathématique par des méthodes principales de l'analyse numérique.

8.1. Equation de Poisson

Nous exposerons à l'aide de certains modèles les concepts à l'origine de la résolution numérique de problèmes relatifs aux équations elliptiques.

8.1.1. Problème de Dirichlet pour l'équation de Poisson en dimension un

Soit le problème de Dirichlet relatif à l'équation de Poisson unidimensionnelle

$$\begin{aligned} -\frac{d^2\varphi}{dx^2} &= f, \qquad 0 < x < 1, \\ \varphi(0) &= a, \qquad \varphi(1) = b, \end{aligned} \tag{1.1}$$

$f(x)$ étant la fonction de sources et a, b des constantes données.

On partage le segment $0 \leqslant x \leqslant 1$ par les points équidistants x_k en N segments partiels $x_k \leqslant x \leqslant x_{k+1}$, l'espacement des points étant $x_k - x_{k-1} = h$. Le problème (1.1) est alors approché à l'ordre 2 en h par

$$\begin{aligned} \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} &= f_k, \qquad k = 1, 2, \ldots, N-1, \\ \varphi_0 &= a, \qquad \varphi_N = b. \end{aligned} \tag{1.2}$$

Selon la loi générale présidant à la formation de schémas aux différences, les points frontières sont éliminés de (1.2) moyennant les conditions aux limites, ce qui change l'opérateur du problème

et on a

$$\frac{2\varphi_1-\varphi_2}{h^2}=\frac{a}{h^2}+f_1,$$

$$\frac{-\varphi_{k-1}+2\varphi_k-\varphi_{k+1}}{h^2}=f_k, \qquad k=2, 3, \ldots, N-2, \tag{1.3}$$

$$\frac{-\varphi_{N-2}+2\varphi_{N-1}}{h^2}=\frac{b}{h^2}+f_{N-1}.$$

On obtient donc un système de $N-1$ équations par rapport aux inconnues $\varphi_1, \varphi_2, \ldots, \varphi_{N-1}$ qui s'écrit sous forme matricielle

$$A\varphi=g, \tag{1.4}$$

où

$$A=\frac{1}{h^2}\begin{Vmatrix} 2 & -1 & 0 & 0\ldots & 0 & 0 \\ -1 & 2 & -1 & 0\ldots & 0 & 0 \\ 0 & -1 & 2 & -1\ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0\ldots & -1 & 2 \end{Vmatrix},$$

$$\varphi=\begin{Vmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{N-1} \end{Vmatrix}, \quad g=\begin{Vmatrix} g_1 \\ g_2 \\ \vdots \\ g_{N-1} \end{Vmatrix},$$

$$g_k=\begin{cases} f_1+\frac{a}{h^2}, & k=1, \\ f_k, & k=2, 3, \ldots, N-2, \\ f_{N-1}+\frac{b}{h^2}, & k=N-1. \end{cases}$$

On vérifie aisément que

$$(A\varphi, \varphi)>0$$

pour $\varphi \neq 0$, i.e. le problème admet une solution unique. On résout (1.3) par la méthode de factorisation:

$$\begin{aligned} &\beta_{k+1}=\frac{1}{2-\beta_k}, \\ &z_{k+1}=\beta_{k+1}(z_k+h^2g_k), \quad k=1, 2, \ldots, N-1, \\ &\varphi_k=\beta_{k+1}\varphi_{k+1}+z_{k+1}, \quad k=N-1, N-2, \ldots, 1. \end{aligned} \tag{1.5}$$

On obtient les valeurs initiales β_2, z_2 en écrivant la troisième équation (1.5) pour $k=1$

$$\varphi_1=\beta_2\varphi_2+z_2$$

et en exigeant qu'elle soit identique à la première équation (1.3) pour β_2, z_2 choisies de façon convenable. On constate facilement qu'on doit poser

$$\beta_2 = \frac{1}{2}, \qquad z_2 = \frac{a + f_1 h^2}{2}. \tag{1.6}$$

On accepte comme conditions initiales pour φ_k

$$\varphi_N = 0. \tag{1.7}$$

8.1.2. Problème de Neumann en dimension un

Soit le problème de Neumann

$$\begin{aligned} -\frac{d^2\varphi}{dx^2} &= f, \\ \frac{d\varphi}{dx} &= a \quad \text{pour} \quad x = 0, \\ \frac{d\varphi}{dx} &= b \quad \text{pour} \quad x = 1, \end{aligned} \tag{1.8}$$

a et b étant des constantes données.

On intègre l'équation (1.8) sur tout le domaine de définition de la solution et on utilise les conditions aux limites, il vient

$$a - b = \int_0^1 f(x)\, dx, \tag{1.9}$$

relation qui est une condition nécessaire de possibilité du problème (1.8). Si $a = b$, cette condition signifie que la somme des sources de substance est égale à 0, i.e. il y a autant de sources positives que de sources négatives.

Ainsi, le problème (1.8) est supposé possible et soit $\varphi_0(x)$ sa solution. On vérifie sans peine que $\varphi(x) = \varphi_0(x) + C$ (C étant une constante arbitraire) en est une autre solution et que toutes les solutions sont définies par cette formule.

On obtient le problème discrétisé d'ordre d'approximation 2 en prolongeant la solution suffisamment régulière au domaine $0 \leqslant x \leqslant 1$ augmenté de h de part et d'autre de la frontière, ce qui fournit le réseau auxiliaire $x_{-1} = -h$, $x_0 = 0$, $x_k = kh$, $x_N = 1$, $x_{N+1} = 1 + h$.

Sur ce réseau on définit une approximation

$$\begin{aligned} \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} &= f_k, \quad k = 0, 1, \ldots, N, \\ \frac{\varphi_1 - \varphi_{-1}}{2h} &= a, \qquad \frac{\varphi_{N+1} - \varphi_{N-1}}{2h} = b. \end{aligned} \tag{1.10}$$

La résolution du problème (1.8) est précédée de l'élimination des conditions aux limites. A cet effet, on élimine dans (1.10) deux nouvelles inconnues φ_{-1}, φ_{N+1}: on résout les relations aux limites (1.10) en φ_{-1}, φ_{N+1}:

$$\varphi_{-1} = \varphi_1 - 2ha, \qquad \varphi_{N+1} = \varphi_{N-1} + 2hb \tag{1.11}$$

et on porte les valeurs obtenues dans les équations aux différences (1.10), il vient

$$\begin{aligned} \frac{\varphi_0 - \varphi_1}{h^2} &= \frac{f_0}{2} - \frac{a}{h}, \\ \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} &= f_k, \qquad k = 1, \ldots, N-1, \\ \frac{-\varphi_{N-1} + \varphi_N}{h^2} &= \frac{f_N}{2} + \frac{b}{h}. \end{aligned} \tag{1.12}$$

On introduit la matrice

$$A = \frac{1}{h^2} \begin{Vmatrix} 1 & -1 & 0 \ldots & 0 & 0 \\ -1 & 2 & -1 \ldots & 0 & 0 \\ 0 & -1 & 2 \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 \ldots & 2 & -1 \\ 0 & 0 & 0 \ldots & -1 & 1 \end{Vmatrix} \tag{1.13}$$

et les vecteurs

$$\varphi = \begin{Vmatrix} \varphi_0 \\ \varphi_1 \\ \varphi_2 \\ \cdots \\ \varphi_N \end{Vmatrix}, \qquad g = \begin{Vmatrix} g_0 \\ g_1 \\ g_2 \\ \cdots \\ g_N \end{Vmatrix},$$

$$g_k = \begin{cases} \dfrac{f_0}{2} - \dfrac{a}{h}, & k = 0, \\ f_k, & k = 1, 2, \ldots, N-1, \\ \dfrac{f_N}{2} + \dfrac{b}{h}, & k = N. \end{cases}$$

Avec ces notations, on a le problème

$$A\varphi = g. \tag{1.14}$$

La première chose à se demander au sujet de A est sa définitivité. Il est connu que la matrice A du problème de Neumann discrétisé est singulière parce qu'il y a nullité de sa plus petite valeur propre. On le constate facilement par recours au problème spectral

$$Au = \lambda u,$$

qui a pour vecteur propre le vecteur u_0 de composantes égales et $\lambda = 0$ pour valeur propre associée. On montre ensuite sans difficulté que $(A\varphi, \varphi) \geqslant 0$. Nous avons étudié à fond ce problème d'algèbre linéaire (voir n° 4.5) et établi une condition nécessaire et suffisante de possibilité. Dans notre cas, cette condition signifie l'orthogonalité de g à u_0, i.e.

$$\sum_{k=0}^{N} g_k = 0. \tag{1.15}$$

Le premier membre s'écrit

$$\sum_{k=0}^{N} g_k = \frac{1}{h}\left[\frac{h}{2}(f_0 + 2f_1 + \ldots + 2f_{N-1} + f_N) - a + b\right]. \tag{1.16}$$

Mais $\frac{h}{2}(f_0 + 2f_1 + \ldots + 2f_{N-1} + f_N)$ n'est rien d'autre que la formule de quadrature des trapèzes pour l'intégrale $\int_0^1 f(x)\,dx$ et son erreur fait que le crochet de (1.16) est ou bien nul ou bien proche de 0. Pour que (1.14) concorde avec la condition de possibilité (1.15), on corrige le « résidu » éventuel et on choisit en qualité de g_k par exemple $\widetilde{g}_k = g_k - \widetilde{g}$, avec $\widetilde{g} = \frac{1}{N+1}\sum_{k=0}^{N} g_k$. On a maintenant

$$\sum_{k=0}^{N} \widetilde{g}_k = 0. \tag{1.17}$$

On a décrit au numéro 4.5 une méthode itérative pour le système compatible

$$A\varphi = \widetilde{g}. \tag{1.18}$$

Le système (1.18) à matrice tridiagonale se prête également à la factorisation à condition de modifier légèrement l'algorithme (A est dégénérée). Le problème modèle considéré permet de bien analyser la méthode et de voir les points cruciaux de l'algorithme.

Considérons donc les formules de factorisation. Dans le cas présent,

$$b_0 = c_0 = b_N = a_N = 1,$$

$$b_k = 2, \quad a_k = c_k = 1, \qquad k = 1, 2, \ldots, N-1,$$

donc

$$\beta_0 = 0, \quad \beta_1 = \beta_2 = \ldots = \beta_N = 1,$$

$$z_0 = 0, \quad z_{k+1} = z_k + \widetilde{g}_k, \qquad k = 0, 1, \ldots, N-1,$$

et de la dernière égalité on déduit facilement $z_{m+1} = \sum_{k=0}^{m} \tilde{g}_k$, en particulier, $z_N = \sum_{k=0}^{N-1} \tilde{g}_k$. Enfin,

$$z_{N+1} = \frac{\tilde{g}_N + z_N}{b_N - a_N \beta_N} = \frac{\tilde{g}_N + \sum_{k=0}^{N-1} \tilde{g}_k}{b_N - a_N \beta_N} = \frac{\sum_{k=0}^{N} \tilde{g}_k}{1-1} = \frac{0}{0}.$$

Ainsi, l'algorithme de factorisation recèle l'indétermination. La chose est naturelle car $\varphi_N = z_{N+1}$ peut être quelconque par suite de la non-unicité de la solution. Ainsi, on prend pour φ_N un nombre arbitraire et on obtient la solution cherchée. On en tiendra certes compte dans le problème de Neumann à un opérateur plus complexe que $-d^2/dx^2$.

8.1.3. Equation de Poisson à deux variables

Dans le cas de plusieurs variables l'approximation constitue un problème non trivial. On illustre au mieux ses difficultés sur un problème en dimension 2 pour l'équation de Poisson. Si la frontière du domaine de définition de la solution est régulière et les fonctions définies sur la frontière possèdent un degré de régularité suffisant, alors une approximation d'ordre 2 de l'équation et des conditions aux limites garantit de règle une solution exacte à l'ordre 2.

Mais si la frontière ou les fonctions ne possèdent pas la propriété voulue, la solution est affectée d'erreurs importantes au voisinage des points singuliers. En effet, si la solution est définie dans un rectangle ou un domaine en L (voir fig. 11), alors il est classique que dans les problèmes à opérateurs elliptiques on voit souvent apparaître au voisinage des points anguleux des singularités logarithmiques ou fractionnaires. Aussi le caractère régulier du réseau et le fait que le problème est approché à l'ordre 2 en les deux variables à l'intérieur du domaine ne garantissent pas une solution exacte à l'ordre 2. Dans ce cas, ou bien on resserre le réseau au voisinage des singularités, ou bien on isole les singularités pour résoudre numériquement le problème régulier de façon à garantir une précision d'ordre 2. On note que s'il y a une concordance déterminée des conditions aux limites et du second membre de l'équation, la solution peut être régulière aux points singuliers de la frontière, ce qui permet d'échapper auxdites difficultés (voir p. ex. Godounov [II]). Nous en avons parlé au Chapitre 6 et on en tiendra nécessairement compte lorsqu'on traitera les problèmes de la physique mathématique par les algorithmes numériques. Nous nous intéresserons donc aux seules méthodes de résolution des équations aux différences intervenant dans les problèmes à plusieurs variables.

Soit l'équation de Poisson de dimension 2

$$-\left(\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial x^2}\right)=f \quad \text{dans} \quad D,$$
$$\varphi=g \quad \text{sur} \quad \partial D, \tag{1.19}$$

le domaine de définition de la solution étant le carré $D = \{0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1\}$. On estime qu'il y a compatibilité mentionnée et que la solution φ présente une régularité suffisante.

Dans D on prend un réseau dont les nœuds (x_k, y_l) sont formés par l'intersection des lignes de coordonnées $x = x_k$ et $y = y_l$. On a l'équation aux différences

$$\frac{-\varphi_{k-1,\,l}-\varphi_{k+1,\,l}-\varphi_{k,\,l-1}-\varphi_{k,\,l+1}+4\varphi_{k,\,l}}{h^2}=f_{k,\,l},$$
$$\varphi_{0,\,l}=a_l, \quad \varphi_{N,\,l}=b_l, \tag{1.20}$$
$$\varphi_{k,\,0}=c_k, \quad \varphi_{k,\,N}=d_k,$$
$$k, l=1, 2, \ldots, N-1,$$

où

$$a_l=g_{0,\,l}, \quad b_l=g_{N,\,l}, \quad c_k=g_{k,\,0}, \quad d_k=g_{k,\,N}$$

(ici $x_k = kh$, $y_l = lh$).

Dans (1.20) on supprime les conditions aux limites et on met le résultat sous forme matricielle. Pour cela, on introduit la matrice A et les vecteurs $\{\varphi_l;\ g_l\}_{l=1}^{N-1}$ par les formules

$$A=\frac{1}{h^2}\begin{Vmatrix} 2 & -1 & 0 & 0\ldots & 0 & 0 \\ -1 & 2 & -1 & 0\ldots & 0 & 0 \\ 0 & -1 & 2 & -1\ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0\ldots & -1 & 2 \end{Vmatrix},$$

$$=\begin{Vmatrix} \varphi_{1l} \\ \varphi_{2l} \\ \varphi_{3l} \\ \vdots \\ \varphi_{N-1,\,l} \end{Vmatrix}, \qquad g_l=\begin{Vmatrix} f_{1l}+\frac{a_l}{h^2} \\ f_{2l} \\ f_{3l} \\ \vdots \\ f_{N-1,\,l}+\frac{b_l}{h^2} \end{Vmatrix}.$$

Avec la matrice B définie par

$$B = h^2A + 2E,$$

le système (1.20) se récrit

$$\frac{1}{h^2}(-\varphi_{l-1}+B\varphi_l-\varphi_{l+1})=g_l, \qquad l=1, 2, \ldots, N-1, \tag{1.21}$$

avec la condition

$$\varphi_0 = c, \qquad \varphi_N = d, \tag{1.22}$$

où

$$c = \begin{Vmatrix} c_1 \\ c_2 \\ \vdots \\ c_{N-1} \end{Vmatrix}, \qquad d = \begin{Vmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N-1} \end{Vmatrix}.$$

Dans les équations (1.21), (1.22) on élimine les conditions aux limites (1.22). On a

$$\begin{gathered} \frac{1}{h^2}(B\varphi_1 - \varphi_2) = g_1 + \frac{1}{h^2} c, \\ \frac{1}{h^2}(-\varphi_{l-1} + B\varphi_l - \varphi_{l+1}) = g_l, \qquad l = 2, 3, \ldots, N-2, \\ \frac{1}{h^2}(-\varphi_{N-2} + B\varphi_{N-1}) = g_{N-1} + \frac{1}{h^2} d, \end{gathered} \tag{1.23}$$

qui devient

$$\Lambda\varphi = F \tag{1.24}$$

à l'aide des vecteurs et des matrices par blocs

$$\Lambda = \frac{1}{h^2} \begin{Vmatrix} B & -E & 0 & 0 \ldots & 0 & 0 \\ -E & B & -E & 0 \ldots & 0 & 0 \\ 0 & -E & B & -E \ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 \ldots & -E & B \end{Vmatrix},$$

$$\varphi = \begin{Vmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{N-1} \end{Vmatrix}, \qquad F = \begin{Vmatrix} \frac{1}{h^2} c + g_1 \\ g_2 \\ \vdots \\ \frac{1}{h^2} d + g_{N-1} \end{Vmatrix},$$

E étant la matrice unité. La matrice Λ se décompose en la somme $\Lambda = \Lambda_1 + \Lambda_2$,

$$\Lambda_1 = \begin{Vmatrix} A & 0 & 0 \ldots & 0 \\ 0 & A & 0 \ldots & 0 \\ 0 & 0 & A \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \ldots & A \end{Vmatrix}, \qquad \Lambda_2 = \frac{1}{h^2} \begin{Vmatrix} 2E & -E & 0 & 0 \ldots & 0 \\ -E & 2E & -E & 0 \ldots & 0 \\ 0 & -E & 2E & -E \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 \ldots & 2E \end{Vmatrix}.$$

On vérifie aisément les égalités

$$\Lambda_1\varphi = \begin{Vmatrix} A\varphi_1 \\ A\varphi_2 \\ \vdots \\ A\varphi_{N-1} \end{Vmatrix}, \qquad \Lambda_2\varphi = \frac{1}{h^2}\begin{Vmatrix} 2\varphi_1 - \varphi_2 \\ \vdots \\ -\varphi_{l-1} + 2\varphi_l - \varphi_{l+1} \\ \vdots \\ -\varphi_{N-2} + 2\varphi_{N-1} \end{Vmatrix}. \tag{1.25}$$

On introduit les notations

$$(\Lambda_1\varphi)_l = A\varphi_l, \qquad l = 1, 2, \ldots, N-1;$$

$$(\Lambda_2\varphi)_k = \begin{cases} \frac{1}{h^2}(2\varphi_1 - \varphi_2)_k, & l = 1, \\ \frac{1}{h^2}(-\varphi_{l-1} + 2\varphi_l - \varphi_{l+1})_k, & l = 2, 3, \ldots, N-2, \\ \frac{1}{h^2}(-\varphi_{N-2} + 2\varphi_{N-1})_k, & l = N-1. \end{cases} \tag{1.26}$$

Il apparaît clairement que $(\Lambda_1\varphi)_l$ et $(\Lambda_2\varphi)_k$ sont les composantes des vecteurs respectifs $\Lambda_1\varphi$ et $\Lambda_2\varphi$ et qu'il est commode de passer aux composantes et d'écrire

$$(\Lambda_1\varphi)_l = \frac{1}{h^2}\begin{Vmatrix} 2\varphi_{1,l} - \varphi_{2,l} \\ \vdots \\ -\varphi_{k-1,l} + 2\varphi_{k,l} - \varphi_{k+1,l} \\ \vdots \\ -\varphi_{N-2,l} + 2\varphi_{N-1,l} \end{Vmatrix}, \quad l = 2, \ldots, N-2,$$

$$(\Lambda_2\varphi)_k = \frac{1}{h^2}\begin{Vmatrix} 2\varphi_{k,l} - \varphi_{k,2} \\ \vdots \\ -\varphi_{k,l-1} + 2\varphi_{k,l} - \varphi_{k,l+1} \\ \vdots \\ -\varphi_{k,N-2} + 2\varphi_{k,N-1} \end{Vmatrix}, \quad \begin{matrix} l = 1, 2, \ldots, N-1; \\ k = 1, 2, \ldots, N-1. \end{matrix}$$

On écrit de même pour le vecteur F:

$$F = \begin{Vmatrix} F_1 \\ F_2 \\ \vdots \\ F_{N-1} \end{Vmatrix},$$

$$F_1 = \begin{Vmatrix} f_{11} + \frac{c_1}{h^2} + \frac{a_1}{h^2} \\ f_{21} + \frac{c_2}{h^2} \\ f_{31} + \frac{c_3}{h^2} \\ \vdots \\ f_{N-1,1} + \frac{c_{N-1}}{h^2} + \frac{b_1}{h^2} \end{Vmatrix}, \qquad F_l = \begin{Vmatrix} f_{1l} + \frac{a_l}{h^2} \\ f_{2l} \\ f_{3l} \\ \vdots \\ f_{N-1,l} + \frac{b_1}{h^2} \end{Vmatrix},$$

$$l = 2, \ldots, N-2,$$

$$F_{N-1}=\left\|\begin{array}{l} f_{1,\,N-1}+\dfrac{d_1}{h^2}+\dfrac{a_{N-1}}{h^2} \\ f_{2,\,N-1}+\dfrac{d_2}{h^2} \\ f_{3,\,N-1}+\dfrac{d_3}{h^2} \\ \vdots \\ f_{N-1,\,N-1}+\dfrac{d_{N-1}}{h^2}+\dfrac{b_{N-1}}{h^2} \end{array}\right\|$$

et on a en passant aux composantes

$$(\Lambda_1\varphi)_l + (\Lambda_2\varphi)_l = F_l. \tag{1.27}$$

Si l'on impose à l'indice l de (1.27) de parcourir toutes les valeurs $l = 1, 2, \ldots, N-1$, on écrit (1.20) sous forme matricielle définitive

$$(\Lambda_1 + \Lambda_2)\,\varphi = F. \tag{1.28}$$

Chaque composante de (1.28) correspond, on le constate sans peine, à l'équation aux différences (1.20) qui tient compte des valeurs données que la solution prend sur la frontière.

Avant d'écrire l'algorithme on cherche les bornes du spectre de l'opérateur Λ.

Les fonctions propres de la matrice Λ étant de la forme (voir point 1.1.4)

$$u_{mp}^{kl} = \sin m\pi kh \sin p\pi lh, \qquad k, l, m, p = 1, 2, \ldots, N-1, \tag{1.29}$$

on a

$$\lambda_{mp} = \frac{4}{h^2}\left(\sin^2\frac{m\pi h}{2} + \sin^2\frac{p\pi h}{2}\right). \tag{1.30}$$

D'où

$$\alpha = \frac{8}{h^2}\sin^2\frac{\pi h}{2}, \qquad \beta = \frac{8}{h^2}\cos^2\frac{\pi h}{2}, \tag{1.31}$$

avec

$$\alpha = \alpha(\Lambda), \qquad \beta = \beta(\Lambda).$$

On aborde l'équation (1.28) moyennant l'une des méthodes itératives du Chapitre 4, à savoir

$$\varphi^{j+1} = \varphi^j - \tau_j(\Lambda\varphi^j - F). \tag{1.32}$$

Le processus itératif s'écrit

$$\varphi^{j+1} = \varphi^j - \tau_j\xi^j,$$
$$\xi^j = \Lambda\varphi^j - F.$$

On continue les itérations (1.32) jusqu'à ce qu'on ait

$$\| \varphi^j - \varphi \| \leqslant \varepsilon,$$

avec ε une constante à priori. Cette estimation a lieu si

$$\| \xi^j \| \leqslant \alpha (\Lambda)\, \varepsilon. \tag{1.33}$$

On accélère en général la convergence si l'on utilise non pas le processus itératif (1.32), mais la méthode des approximations successives avec accélération au sens de Tchebycheff

$$\varphi^{j+1} = \varphi^j - \tau_j B^{-1} (\Lambda \varphi^j - F), \tag{1.34}$$

où

$$B = (E + \sigma\Lambda_1)(E + \sigma\Lambda_2);$$
$$\sigma = \frac{2}{\sqrt{\alpha\beta}}; \quad \alpha = \alpha(\Lambda); \quad \beta = \beta(\Lambda).$$

On localise le spectre de $B^{-1}\Lambda$. Comme les matrices Λ_1 et Λ_2 possèdent une base en commun, il y a entre les valeurs propres λ_1 de Λ_1 et λ_2 de Λ_2, d'une part, et les valeurs propres du problème (voir point 1.1.3) de l'autre

$$B^{-1}\Lambda u = \lambda (B^{-1}\Lambda)\, u,$$

la relation

$$\lambda (B^{-1}\Lambda) = \frac{\lambda_1 + \lambda_2}{\left(1 + \frac{2\lambda_1}{\sqrt{\alpha\beta}}\right)\left(1 + \frac{2\lambda_2}{\sqrt{\alpha\beta}}\right)}, \tag{1.35}$$
$$\alpha/2 \leqslant \lambda_1 \leqslant \beta/2, \quad \alpha/2 \leqslant \lambda_2 \leqslant \beta/2,$$

qui se ramène à

$$\lambda (B^{-1}\Lambda) = \frac{\sqrt{\alpha\beta}}{2} f(x, y), \tag{1.36}$$

avec

$$x = \frac{2\lambda_1}{\sqrt{\alpha\beta}}, \quad y = \frac{2\lambda_2}{\sqrt{\alpha\beta}}; \quad f(x, y) = \frac{x + y}{(1+x)(1+y)}. \tag{1.37}$$

Ainsi, une condition suffisante pour qu'on localise le spectre de la matrice $B^{-1}\Lambda$ est qu'on trouve les valeurs extrêmes de la fonction $f(x, y)$ dans le carré $\sqrt{\frac{\alpha}{\beta}} \leqslant x,\ y \leqslant \sqrt{\frac{\beta}{\alpha}}$. Les dérivées de cette fonction étant

$$\frac{\partial f}{\partial x} = \frac{1-y}{1+y} \frac{1}{(1+x)^2}, \quad \frac{\partial f}{\partial y} = \frac{1-x}{1+x} \frac{1}{(1+y)^2}, \tag{1.38}$$

on montre aisément que $f(x, y)$ est maximale en deux points anguleux :

$$\max_{\sqrt{\alpha/\beta}\leqslant x,\ y\leqslant\sqrt{\beta/\alpha}} f(x, y) = f\left(\sqrt{\frac{\alpha}{\beta}},\ \sqrt{\frac{\beta}{\alpha}}\right) = f\left(\sqrt{\frac{\beta}{\alpha}},\ \sqrt{\frac{\alpha}{\beta}}\right), \quad (1.39)$$

et qu'elle est minimale en deux points restants :

$$\min_{\sqrt{\alpha/\beta}\leqslant x,\ y\leqslant\sqrt{\beta/\alpha}} f(x, y) = f\left(\sqrt{\frac{\alpha}{\beta}},\ \sqrt{\frac{\alpha}{\beta}}\right) = f\left(\sqrt{\frac{\beta}{\alpha}},\ \sqrt{\frac{\beta}{\alpha}}\right). \quad (1.39a)$$

Il en résulte compte tenu de (1.36)

$$\alpha(B^{-1}\Lambda) = \frac{\alpha}{(1+\sqrt{\alpha/\beta})^2},\quad \beta(B^{-1}\Lambda) = \frac{1}{2}\sqrt{\frac{\alpha}{\beta}}\,\frac{(\alpha+\beta)}{(1+\sqrt{\alpha/\beta})^2}, \quad (1.40)$$

autrement dit, on a asymptotiquement pour $\beta(\Lambda) \gg \alpha(\Lambda)$

$$p(B^{-1}\Lambda) = \frac{\beta(B^{-1}\Lambda)}{\alpha(B^{-1}\Lambda)} = \frac{1}{2}\frac{\alpha+\beta}{\sqrt{\alpha\beta}} \cong \frac{1}{2}\sqrt{\frac{\beta}{\alpha}} = \frac{1}{2}[p(\Lambda)]^{1/2}, \quad (1.41)$$

$$s = \frac{2}{\sqrt{p(B^{-1}\Lambda)}} = 2^{3/2}p^{-1/4}(\Lambda). \quad (1.42)$$

Les itérations (1.34) procèdent comme suit :

$$\begin{aligned} \xi^j &= \Lambda\varphi^j - F, \\ (E+\sigma\Lambda_1)\,\xi^{j+1/2} &= \xi^j, \\ (E+\sigma\Lambda_2)\,\xi^{j+1} &= \xi^{j+1/2}, \\ \varphi^{j+1} &= \varphi^j - \tau_j\xi^{j+1}, \end{aligned} \quad (1.43)$$

τ_j étant cherché par la méthode d'optimisation choisie.

En passant aux composantes, l'algorithme pour la deuxième et la troisième équation (1.43) s'écrit : on résout d'abord le problème pour $l = 1, 2, \ldots, N-1$

$$\begin{aligned} (1+2\sigma)\,\xi^{j+1/2}_{1,l} - \sigma\xi^{j+1/2}_{2,l} &= \xi^j_{1,l}, \\ -\sigma\xi^{j+1/2}_{k-1,l} + (1+2\sigma)\,\xi^{j+1/2}_{k,l} - \sigma\xi^{j+1/2}_{k+1,l} &= \xi^j_{k,l}, \\ -\sigma\xi^{j+1/2}_{N-2,l} + (1+2\sigma)\,\xi^{j+1/2}_{N-1,l} &= \xi^j_{N-1,l}, \end{aligned} \quad (1.44)$$

puis pour $k=1, 2, \ldots, N-1$

$$\begin{aligned}(1+2\sigma)\,\xi^{j+1}_{k,1}-\sigma\xi^{j+1}_{k,2}&=\xi^{j+1/2}_{k,l},\\ -\sigma\xi^{j+1}_{k,l-1}+(2\sigma+1)\,\xi^{j+1}_{k,l}-\sigma\xi^{j+1}_{k,l-1}&=\xi^{j+1/2}_{k,l},\\ -\sigma\xi^{j+1}_{k,N-2}+(1+2\sigma)\,\xi^{j+1}_{k,N-1}&=\xi^{j+1/2}_{k,N-1}.\end{aligned} \tag{1.45}$$

Les équations (1.44) et (1.45) sont traitées par factorisation.

S'agissant du problème de Neumann en dimension 2, on aboutit encore à (1.28) à la différence que les composantes à définir ne sont plus $(N-1)^2$ (cas de Dirichlet), mais $(N+1)^2$.

Dans le problème de Neumann, $\alpha(\Lambda)=0$. C'est une différence essentielle avec le schéma de mise en œuvre du problème de Dirichlet discrétisé. Par conséquent, le second membre F et la solution approchée φ^j sont orthogonaux à chaque pas à un vecteur de composantes identiques, i.e. à l'entrée de chaque pas d'itération on retranche de toutes les composantes de ξ^j la constante $\frac{1}{(N+1)^2}\sum\limits_{k,l}\xi^j_{k,l}$.

Dans ce procédé d'orthogonalisation, l'espace de Hilbert primitif devient par élimination des vecteurs de composantes identiques le sous-espace Φ dont les éléments sont pris pour vecteurs d'essai, si bien qu'on assimile la borne inférieure du spectre à la plus petite valeur propre non nulle. Il est connu que le spectre du problème de Neumann relatif au laplacien discrétisé se définit par la formule

$$\lambda_{mp}=\frac{4}{h^2}\left(\sin^2\frac{m\pi h}{2}+\sin^2\frac{p\pi h}{2}\right), \qquad m,\ p=0, 1, \ldots, N, \tag{1.46}$$

et qu'on a pour les bornes

$$\alpha^*(\Lambda)=\frac{8}{h^2}\sin^2\frac{\pi h}{2}, \qquad \beta(\Lambda)=\frac{8}{h^2}. \tag{1.47}$$

Il y a intérêt à choisir un paramètre σ tel que

$$\sigma=\sqrt{\frac{2}{\alpha^*\beta}} \tag{1.48}$$

et à accepter pour critère d'arrêt du processus itératif (1.33)

$$\|\xi^j\|<\alpha^*(\Lambda)\,\varepsilon. \tag{1.49}$$

Il n'existe pas d'autre différence entre l'algorithme numérique pour le problème de Neumann attaché à l'équation de Poisson et l'algorithme considéré pour le cas de Dirichlet.

8.1.4. Problème de conditions aux limites

Les méthodes considérées pour l'équation de Poisson autorisent une conclusion importante et assez générale sur l'élaboration d'algorithmes efficaces pour les problèmes aux limites de la physique

mathématique. Cela concerne en premier lieu le problème de conditions aux limites. Plus haut on s'est tenu à éliminer les conditions aux limites, ces contraintes supplémentaires imposées à la solution, et à les modifier compte tenu des problèmes aux différences associés. Cette tactique a sa raison d'être : les conditions aux limites étant « supprimées », on construit les algorithmes de calcul sans avoir à les vérifier du moment que les équations aux différences modifiées en tiennent compte automatiquement. La chose est essentielle pour tous les problèmes stationnaires ou non stationnaires, surtout pour ces derniers qui exigent une analyse sérieuse du problème de conditions aux limites. Aussi, nous nous sommes abstenu au Chapitre 5 de décomposer les problèmes non stationnaires en des problèmes différentiels simples, sinon on aurait à s'occuper des conditions aux limites compatibles avec le système décomposé. Selon nous, il vaut mieux associer au problème initial un système d'équations discrétisées en les variables d'espace et en éliminer les valeurs aux limites des fonctions à l'aide des conditions aux limites discrétisées qui sont compatibles sur le plan précision avec l'équation aux différences. Ce n'est qu'après qu'on approche par rapport au temps par les méthodes de décomposition ou autres. Cet artifice évite l'étude de la compatibilité des conditions aux limites à chaque étape de la formation des algorithmes de décomposition.

Nous voulons attirer l'attention du lecteur sur une circonstance intimement liée avec le problème de conditions aux limites. Dans certains problèmes de la physique mathématique traités par les schémas aux différences, on a intérêt à représenter la solution par la série de Fourier suivant les éléments propres de l'opérateur du problème discrétisé. Nous y avons recouru à maintes reprises pour tirer au clair les propriétés des algorithmes de calcul. La méthode en question ne joue cependant que si le problème discrétisé présente des conditions aux limites homogènes. Dans le cas contraire, on transforme au préalable le problème de façon que les conditions deviennent homogènes. Effectuons cette transformation sur le problème simple (1.1). On considère le système aux différences (1.2), on adjoint au domaine de définition $\{x_k\}$, $k = 1, 2, \ldots, N-1$, de la solution deux nœuds $x = 0$ $(k = 0)$ et $x = 1$ $(k = 1)$ et on prolonge la solution à ces points par

$$\varphi_0 = 0, \quad \varphi_N = 0.$$

Cette opération purement formelle n'a certes rien à faire avec la valeur réelle de la solution en $x = 0$ et $x = 1$ définis par les conditions aux limites non homogènes (1.1). Cela signifie qu'une fois le problème résolu, on chasse du domaine de définition les points $k = 0$ et $k = N$ où la solution est définie par (1.1). Avec ce procédé,

on substitue à (1.2) le problème équivalent

$$\frac{-\varphi_{k-1}+2\varphi_k-\varphi_{k-1}}{h^2}=g_k, \qquad k=1,\ 2,\ \ldots,\ N-1, \qquad \varphi_0=0, \quad \varphi_N=0, \tag{1.50}$$

où

$$g_k=\begin{cases} \frac{a}{h^2}+f_1, & k=1, \\ f_k, & k=2,\ 3,\ \ldots,\ N-2, \\ \frac{b}{h^2}+f_{N-1}, & k=N-1. \end{cases} \tag{1.51}$$

Ainsi, le problème (1.1) avec conditions aux limites non homogènes devient par une certaine transformation du second membre des équations aux différences un problème homogène dont on cherche la solution par la méthode de Fourier.

Dans le problème de Neumann non homogène (1.10) on élargit le domaine correspondant par adjonction des nœuds x_{-1} et x_{N+1}, ce qui donne le problème

$$\frac{-\varphi_{k-1}+2\varphi_k-\varphi_{k+1}}{h^2}=g_k, \qquad k=0,\ 1,\ 2,\ \ldots,\ N, \qquad \varphi_{-1}=\varphi_1, \quad \varphi_{N+1}=\varphi_{N-1}, \tag{1.52}$$

où

$$g_k=\begin{cases} f_0-\frac{2a}{h}, & k=0, \\ f_k, & k=1,\ 2,\ \ldots,\ N-1, \\ f_N+\frac{2b}{h}, & k=N. \end{cases} \tag{1.53}$$

On cherche la solution et on en élimine les composantes auxiliaires φ_{-1}, φ_{N+1}.

On traite de même les problèmes à plusieurs variables, stationnaires ou non.

Les systèmes (1.50) à (1.53) sont traités, si besoin est, comme les problèmes non homogènes (1.1) et (1.10) : on élimine des équations les valeurs aux limites des fonctions. On est conduit une fois de plus à (1.3) et à (1.12).

Dans la suite, nous ne considérerons que les conditions aux limites homogènes parce que les algorithmes décrits permettent ou bien de supprimer les conditions qui ne le sont pas, ou bien de les mettre sous forme homogène.

8.2. Equation de la chaleur

Le problème de la chaleur est un problème non stationnaire typique de la physique mathématique. Historiquement, l'équation de la chaleur a été à l'origine de nombreuses idées de base de l'Analyse numérique et d'algorithmes performants pour les problèmes de la physique mathématique. Les recherches dans ce domaine remontent à l'ouvrage classique de O'Brien, Hyman, Kaplan [VII] qui soulève et discute le problème de convergence des solutions approchées. La résolution numérique des équations de la chaleur fait l'objet de plusieurs monographies et articles originaux (dont le livre de Sauliev [III ; 15] fait le tour), si bien que nous nous limiterons, pour notre part, à certaines méthodes qu'on rencontre le plus souvent dans les applications.

8.2.1. Problème de la chaleur en dimension 1

Soit le problème simple portant sur la propagation de la chaleur dans une tige homogène limitée avec sources chauffantes intérieures, dont les extrémités absorbent et abandonnent de la chaleur :

$$\begin{aligned} &\frac{1}{c^2}\frac{\partial \varphi}{\partial t} = \frac{\partial^2 \varphi}{\partial x^2} + f(x, t), \\ &\varphi(0, t) = a(t), \qquad \varphi(1, t) = b(t), \\ &\varphi(x, 0) = \varphi^0(x), \end{aligned} \tag{2.1}$$

où f, a, b et φ^0 sont des fonctions données suffisamment régulières, $c = \text{const}$ et les variables x et t parcourent toutes les valeurs du domaine de définition $D \times D_t = \{0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T\}$ de la solution.

On cherche la solution par une méthode des différences finies. On approche (2.1) à l'ordre 2 en x. Comme dans le problème (1.1), on partage l'intervalle $0 \leqslant x \leqslant 1$ par les points x_k, $k = 1, 2, \ldots$ $\ldots, N-1$, en N intervalles partiels de longueur $h = \frac{1}{N}$. On est conduit à un problème analogue à (1.2), à savoir

$$\begin{aligned} &\frac{1}{c^2}\frac{d\varphi_k}{dt} + \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} = f_k(t), \\ &\varphi_0 = a(t), \qquad \varphi_N = b(t), \\ &\varphi_k = \varphi_k^0 \quad \text{pour} \quad t = 0, \\ &k = 1, 2, \ldots, N-1, \end{aligned} \tag{2.2}$$

qui s'écrit sous forme matricielle

$$\begin{aligned} &\frac{1}{c^2}\frac{d\varphi}{dt} + A\varphi = g, \\ &\varphi = \varphi^0 \quad \text{pour} \quad t = 0, \end{aligned} \tag{2.3}$$

A étant une matrice positive et $\varphi(t)$ et $g(t)$ deux fonctions vectorielles définies pour tout t dans (1.4).

Les raisonnements du point 8.1.4 aidant, le système (2.2) prend la forme

$$\frac{1}{c^2}\frac{d\varphi_k}{dt}+\frac{-\varphi_{k-1}+2\varphi_k-\varphi_{k+1}}{h^2}=g_k,$$

$$\varphi_0=0, \quad \varphi_N=0, \tag{2.4}$$

$$\varphi_k=\varphi_k^0 \quad \text{pour} \quad t=0,$$

g_k étant défini par les formules (1.51).

On connaît dès le point 8.1.4 que la solution du problème (2.4) a le sens aux seuls nœuds $x_1, x_2, \ldots, x_{N-1}$.

Passons au système d'équations différentielles ordinaires (2.4). On cherche sa solution en intégrant chaque équation par rapport au temps sur l'intervalle $t_j \leqslant t \leqslant t_{j+1}$. On a

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}=\frac{\bar{\varphi}_{k-1}^j-2\bar{\varphi}_k^j+\bar{\varphi}_{k+1}^j}{h^2}+\bar{g}_k^j, \tag{2.5}$$

où

$$\bar{\varphi}_k^j=\frac{1}{\Delta t}\int_{t_j}^{t_{j+1}}\varphi_k\,dt, \qquad \bar{g}_k^j=\frac{1}{\Delta t}\int_{t_j}^{t_{j+1}}g_k\,dt \tag{2.6}$$

et

$$\varphi_k^j=\varphi_k(t_j), \qquad \tau=c^2\Delta t.$$

Selon l'approximation (2.6), on aboutit à telle ou telle équation aux différences. Supposons vraie l'une des formules simples

$$\frac{1}{\Delta t}\int_{t_j}^{t_{j+1}}\varphi_k\,dt=\begin{cases}\varphi_k^j,\\ \varphi_k^{j+1},\\ \frac{1}{2}(\varphi_k^j+\varphi_k^{j+1});\end{cases} \qquad \frac{1}{\Delta t}\int_{t_j}^{t_{j+1}}g_k\,dt=\begin{cases}g_k^j,\\ g_k^{j+1},\\ \frac{1}{2}(g_k^{j+1}+g_k^j),\end{cases}$$

auquel cas on aboutit aux schémas aux différences le plus connus:

schéma triangulaire explicite $(\cdot\,\cdot)$

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}=\frac{\varphi_{k-1}^j-2\varphi_k^j+\varphi_{k+1}^j}{h^2}+g_k^j; \tag{2.7}$$

schéma triangulaire implicite $(\cdot\cdot\cdot)$

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}=\frac{\varphi_{k-1}^{j+1}-2\varphi_k^{j+1}+\varphi_{k+1}^{j+1}}{h^2}+g_k^{j+1}; \tag{2.8}$$

schéma de Crank-Nicholson (⋮⋮⋮)

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}=\frac{\varphi_{k-1}^{j+1}-2\varphi_k^{j+1}+\varphi_{k+1}^{j+1}}{2h^2}+\frac{\varphi_{k-1}^j-2\varphi_k^j+\varphi_{k+1}^j}{2h^2}+\frac{g_k^{j+1}+g_k^j}{2}. \tag{2.9}$$

On adjoint aux systèmes (2.7)-(2.9) les conditions aux limites

$$\varphi_0^j=0, \quad \varphi_N^j=0. \tag{2.10}$$

Le schéma triangulaire (⋮⋮) s'explicite par rapport à l'inconnue

$$\varphi^{j+1}=\varphi^j+\mu(\varphi_{k-1}^j-2\varphi_k^j+\varphi_{k+1}^j)+\tau g_k^j, \qquad \varphi_0^j=0, \quad \varphi_N^j=0, \quad \mu=\frac{\tau}{h^2}. \tag{2.11}$$

Le schéma triangulaire implicite (⋮⋮) est de mise en œuvre plus délicate qui se ramène à la résolution de l'équation aux différences

$$-\varphi_{k-1}^{j+1}+\left(2+\frac{1}{\mu}\right)\varphi_k^{j+1}-\varphi_{k+1}^{j+1}=\frac{1}{\mu}\varphi_k^j+h^2g_k^{j+1}, \qquad \varphi_0^{j+1}=0, \quad \varphi_N^{j+1}=0. \tag{2.12}$$

Le schéma de Crank-Nicholson (⋮⋮⋮) s'écrit enfin

$$-\xi_{k-1}^{j+1}+\frac{1+2\mu}{\mu}\xi_k^{j+1}-\xi_{k+1}^{j+1}=\frac{2}{\mu}\varphi_k^j+h^2g_k^{j+1/2}, \qquad \xi_0^{j+1}=0, \quad \xi_N^{j+1}=0, \qquad \varphi_k^{j+1}=2\xi_k^{j+1}-\varphi_k^j, \tag{2.13}$$

où

$$g_k^{j+1/2}=\frac{1}{2}(g_k^{j+1}+g_k^j).$$

Les problèmes (2.12), (2.13) se prêtent bien à la méthode de factorisation.

Voyons la propriété de stabilité des schémas aux différences (2.7)-(2.9) avec les conditions (2.10). On développe la solution en série de Fourier suivant le système complet de fonctions $\{\sin n\pi kh\}$, $h = 1/N$, vérifiant (2.10). Soit

$$\varphi_k^j=\sum_{n=1}^{N-1}\Phi_n^j\sin n\pi kh, \qquad g_k^j=\sum_{n=1}^{N-1}G_n^j\sin n\pi kh, \tag{2.14}$$

où

$$\Phi_n^j=\frac{1}{q_n}\sum_{k=1}^{N-1}\varphi_k^j\sin n\pi kh, \qquad G_n^j=\frac{1}{q_n}\sum_{k=1}^{N-1}g_k^j\sin n\pi kh,$$

$$q_n=\sum_{k=1}^{N-1}\sin^2 n\pi kh.$$

On multiplie par sin $ln\pi h$ le résultat de la substitution de (2.14) dans (2.7)-(2.9), puis on somme par rapport à l. Il vient les relations de récurrence pour les coefficients de Fourier. On a notamment pour le schéma (.:.)

$$\frac{\Phi_n^{j+1}-\Phi_n^j}{\tau}+\lambda_n\Phi_n^j=G_n^j, \tag{2.15}$$

pour (:.:)

$$\frac{\Phi_n^{j+1}-\Phi_n^j}{\tau}+\lambda_n\Phi_n^{j+1}=G_n^{j+1}, \tag{2.16}$$

et dans le cas (:::) M

$$\frac{\Phi_n^{j+1}-\Phi_n^j}{\tau}+\lambda_n\frac{\Phi_n^{j+1}+\Phi_n^j}{2}=\frac{G_n^{j+1}+G_n^j}{2}. \tag{2.17}$$

Ici

$$\lambda_n=\frac{4}{h^2}\sin^2\frac{n\pi h}{2}. \tag{2.18}$$

On résout (2.15) à (2.17):

$$\begin{aligned}\Phi_n^{j+1}&=(1-\tau\lambda_n)\,\Phi_n^j+\tau G_n^j,\\ \Phi_n^{j+1}&=\frac{1}{1+\tau\lambda_n}\Phi_n^j+\frac{\tau}{1+\tau\lambda_n}G_n^{j+1},\\ \Phi_n^{j+1}&=\frac{1-\frac{\tau}{2}\lambda_n}{1+\frac{\tau}{2}\lambda_n}\Phi_n^j+\frac{\tau}{2}\frac{G_n^{j+1}+G_n^j}{1+\frac{\tau}{2}\lambda_n}.\end{aligned} \tag{2.19}$$

On note que

$$\frac{4}{h^2}\sin^2\frac{\pi h}{2}\leqslant\lambda_n\leqslant\frac{1}{h^2}\cos^2\frac{\pi h}{2}<\frac{4}{h^2} \tag{2.20}$$

qui deviennent pour $\frac{\pi h}{2}\ll 1$ les inégalités asymptotiques

$$\pi^2\leqslant\lambda_n\leqslant\frac{4}{h^2}. \tag{2.21}$$

On conclut, compte tenu de (2.20), que si

$$\tau\leqslant\frac{h^2}{2\cos^2\frac{\pi h}{2}}, \tag{2.22}$$

il y a instabilité du schéma triangulaire (.:.) parce que tous les n vérifient l'inégalité

$$|\,1-\tau\lambda_n\,|<1. \tag{2.23}$$

On remplace (2.22) par la condition suffisante

$$\tau\leqslant\frac{h^2}{2}. \tag{2.24}$$

S'agissant du schéma triangulaire implicite et du schéma de Crank-Nicholson, on a les inégalités respectives

$$\left|\frac{1}{1+\tau\lambda_n}\right|<1, \qquad \left|\frac{1-\frac{\tau}{2}\lambda_n}{1+\frac{\tau}{2}\lambda_n}\right|<1$$

quels que soient n et $\tau>0$, i.e. le schéma (. :.) est stable et les deux autres le sont absolument sous la condition (2.22). Dans les calculs on utilise de préférence le schéma de Crank-Nicholson d'ordre d'approximation 2.

Dans le cas de Neumann relatif à l'équation de la chaleur, on définit l'opérateur A et les fonctions vectorielles φ et g à partir du problème (1.14) et on aboutit à des équations analogues à (2.5), (2.6), dont la résolution ne présente pas de difficulté. Il y a lieu de dire que tandis que l'harmonique associé à $\lambda = 0$ dans le problème de Neumann pour l'équation de Laplace est une composante parasite à « filtrer » de la solution approchée, il constitue une partie intégrante de la solution dans le problème de la propagation de la chaleur. En effet, cette composante décrit l'augmentation ou la diminution générale de température dues aux sources positives et négatives extérieures à la tige.

8.2.2. Problème de la chaleur en dimension 2

Soit le problème de la chaleur en dimension deux

$$\begin{aligned} &\frac{1}{c^2}\frac{\partial\varphi}{\partial t}-\Delta\varphi=f \quad \text{dans} \quad D\times D_t,\\ &\varphi=g \quad \text{ou} \quad \frac{\partial\varphi}{\partial n}=g \quad \text{sur} \quad \partial D\times D_t,\\ &\varphi=s\,(x,\,y) \quad \text{dans} \quad D \quad \text{pour} \quad t=0, \end{aligned} \tag{2.25}$$

D étant le carré $\{0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant 1\}$.

On discrétise (2.25) en x et y et on obtient par des transformations analogues à celles du cas de Dirichlet à deux variables :

$$\begin{aligned} &\frac{1}{c^2}\frac{d\varphi}{dt}+(\Lambda_1+\Lambda_2)\varphi=F,\\ &\varphi=s \quad \text{pour} \quad t=0, \end{aligned} \tag{2.26}$$

avec les matrices Λ_1, Λ_2 et les fonctions vectorielles φ et F définies comme dans le problème (1.24) à la différence que les composantes de φ et F dépendent maintenant de la variable t. On note que dans le cas de conditions aux limites de Dirichlet,

$$\Lambda_1>0, \qquad \Lambda_2>0, \tag{2.27}$$

et pour les conditions de Neumann

$$\Lambda_1 \geqslant 0, \quad \Lambda_2 \geqslant 0. \tag{2.28}$$

On partage l'intervalle $0 \leqslant t \leqslant T$ par les points t_j en sous-intervalles et on pose $\tau = c^2 \Delta t$. On cherche à approcher le problème (2.26) par la technique de splitting-up et on a sur $t_{j-1} \leqslant t \leqslant t_{j+1}$

$$\begin{aligned}
\left(E + \tfrac{\tau}{2}\Lambda_1\right) \varphi^{j-1/2} &= \left(E - \tfrac{\tau}{2}\Lambda_1\right) \varphi^{j-1}, \\
\left(E + \tfrac{\tau}{2}\Lambda_2\right) (\varphi^j - \tau f^j) &= \left(E - \tfrac{\tau}{2}\Lambda_2\right) \varphi^{j-1/2}, \\
\left(E + \tfrac{\tau}{2}\Lambda_2\right) \varphi^{j+1/2} &= \left(E - \tfrac{\tau}{2}\Lambda_2\right) (\varphi^j + \tau f^j), \\
\left(E + \tfrac{\tau}{2}\Lambda_1\right) \varphi^{j+1} &= \left(E - \tfrac{\tau}{2}\Lambda_1\right) \varphi^{j+1/2}.
\end{aligned} \tag{2.29}$$

On note qu'en passant aux coordonnées dans le second membre on utilise des schémas explicites. Il s'agit donc d'un système d'équations sous forme de coordonnées qui s'écrivent comme des schémas à trois points analogues à (1.43) où le rôle de σ revient au paramètre $\tau/2$ et celui de ξ_k^j aux composantes des seconds membres de (2.29).

On procède de même pour les conditions aux limites de Neumann.

8.3. Equation des vibrations

On rencontre fréquemment dans les applications les équations hyperboliques justiciables des méthodes numériques bien étudiées. Ces équations ont cela de particulier que le domaine de dépendance de leur solution est limité par le cône caractéristique, si bien que le domaine $D \times D_t$ extérieur à ce cône n'a aucune influence sur la solution au point considéré. Leur autre particularité est que les solutions du problème de départ sont telles qu'il existe en général des solutions non régulières, circonstance à retenir lorsqu'on se propose d'élaborer des schémas de calcul.

Considérons le problème unidimensionnel portant sur les petites vibrations d'une corde homogène :

$$\begin{aligned}
&\frac{1}{c^2}\frac{\partial^2 \varphi}{\partial t^2} = \frac{\partial^2 \varphi}{\partial x^2} + f(x, t), \\
&\varphi(0, t) = a(t), \quad \varphi(1, t) = b(t), \\
&\varphi(x, 0) = p(x), \quad \frac{\partial \varphi}{\partial t}(x, 0) = q(x),
\end{aligned} \tag{3.1}$$

c étant la vitesse de propagation des perturbations le long de la corde et $a(t)$, $b(t)$, $f(x, t)$, $p(x)$, $q(x)$ des fonctions données.

La théorie des cordes vibrantes n'a guère de lacunes. Notre but à nous est de discuter la résolution numérique de (3.1). On com-

mence par ramener ce problème par discrétisation en x à un système d'équations différentielles ordinaires par rapport au temps. Moyennant un procédé du n° 8.2, on approche (3.1) par

$$\frac{1}{c^2}\frac{d^2\varphi_k}{dt^2}+\frac{-\varphi_{k-1}+2\varphi_k-\varphi_{k+1}}{h^2}=g_k(t),$$
$$\varphi_0=0, \qquad \varphi_N=0, \tag{3.2}$$
$$\varphi_k=p_k, \qquad \frac{d\varphi_k}{dt}=q_k \quad \text{pour} \quad t=0,$$

$g_k(t)$ étant la fonction définie par (1.15) (comme pour (2.4)).

Il est déjà apparu que la solution $\varphi_k(t)$ de (3.2) n'a de sens qu'aux nœuds $k=1, 2, \ldots, N-1$. Elle est définie aux points $k=0$ et $k=N$ par les conditions aux limites de (3.1), et les conditions homogènes (3.2) proviennent d'une extension spéciale du domaine de définition de la solution et elles ne sont pas des approximations de $\varphi(x, t)$ en x_0 et x_N. On note que le problème (3.2) est approché d'ordre 2 en $h=\Delta x$.

Pour discrétiser (3.2) par rapport au temps, on établit un système de nœuds $t=t_j$ (tel que $t_{j+1}=t_j+\Delta t$) et on considère les schémas les plus usités, à savoir

le schéma de la croix explicite (.:.)

$$\frac{\varphi_k^{j+1}-2\varphi_k^j+\varphi_k^{j-1}}{\tau^2}=\frac{\varphi_{k-1}^j-2\varphi_k^j+\varphi_{k+1}^j}{h^2}+g_k^j,$$
$$\varphi_0^j=0, \quad \varphi_N^j=0, \quad \varphi_k^0=p_k, \tag{3.3}$$
$$\varphi_k^1=p_k+\Delta t q_k+\frac{\tau^2}{2}\left(\frac{p_{k-1}-2p_k+p_{k+1}}{h^2}+g_k^0\right),$$

où $\tau=c\,\Delta t$, et le schéma implicite du type de Crank-Nicholson (⋮⋮⋮)

$$\frac{\varphi_k^{j+1}-2\varphi_k^j+\varphi_k^{j-1}}{\tau^2}=$$
$$=\frac{\varphi_{k-1}^{j+1}-2\varphi_k^{j+1}+\varphi_{k+1}^{j+1}}{2h^2}+\frac{\varphi_{k-1}^{j-1}-2\varphi_k^{j-1}+\varphi_{k+1}^{j-1}}{2h^2}+g_k^j, \tag{3.4}$$
$$\varphi_0^{j+1}=0, \quad \varphi_N^{j+1}=0, \quad \varphi_k^0=p_k,$$
$$\varphi_k^1=\varphi_k^0+\Delta t q_k+\frac{\tau^2}{2}\left(\frac{p_{k-1}-2p_k+p_{k+1}}{h^2}+g_k^0\right).$$

On développe en série de Taylor et on établit que les schémas aux différences (3.3) et (3.4) approchent, pour $\varphi(x, t)$ suffisamment régulières, le problème primitif à l'ordre 2 en h et τ.

On cherche la solution des problèmes (3.3), (3.4) en développant en série de Fourier par rapport au système de fonctions propres

$$u_n(k)=\sin n\pi kh, \qquad k, n=1, 2, \ldots, N-1$$

(on imite donc le cas (2.14)). On trouve les relations de récurrence suivantes pour les coefficients de Fourier:

schéma de la croix explicite:

$$\frac{\Phi_n^{j+1}-2\Phi_n^j+\Phi_n^{j-1}}{\tau^2}+\frac{4}{h^2}\sin^2\frac{n\pi h}{2}\,\Phi_n^j=G_n^j,$$

$$\Phi_n^0=P_n, \tag{3.5}$$

$$\Phi_n^1=P_n+\Delta t Q_n+\frac{2\tau^2}{h^2}\sin^2\frac{n\pi h}{2}\,P_n+\frac{\tau^2}{2}\,G_n^0,$$

avec Φ_n, G_n, P_n et Q_n les coefficients de Fourier des φ_k, g_k, p_k et q_k respectivement;

schéma implicite $\left(\begin{smallmatrix}\cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot\end{smallmatrix}\right)$:

$$\frac{\Phi_n^{j+1}-2\Phi_n^j+\Phi_n^{j-1}}{\tau^2}+\frac{4}{h^2}\sin^2\frac{n\pi h}{2}\,\frac{\Phi_n^{j+1}+\Phi_n^{j-1}}{2}=G_n^j,$$

$$\Phi_n^0=P_n, \tag{3.6}$$

$$\Phi_n^1=P_n+\Delta t Q_n+\frac{2\tau^2}{h^2}\sin^2\frac{n\pi h}{2}\,P_n+\frac{\tau^2}{2}\,G_n^0.$$

Quel est le critère de stabilité numérique pour les problèmes à données initiales régulières? Pour répondre à cette question on étudie le comportement des solutions linéairement indépendantes de (3.5) et (3.6) homogènes en fonction de l'indice j. La solution est cherchée sous la forme

$$\Phi_n^j=A_n\eta_n^j, \tag{3.7}$$

avec A_n et η_n des constantes, le dernier j en indice supérieur étant un exposant. On pose $G_n^j=0$ dans (3.5), (3.6) et on y substitue (3.7), il vient les expressions suivantes pour η_n:

$$\eta_n^2-2(1-\mu_n^2)\,\eta_n+1=0 \tag{3.8}$$

(schéma de la croix $\left(\begin{smallmatrix}\cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot\end{smallmatrix}\right)$) et

$$\eta_n^2-\frac{2}{1+\mu_n^2}\,\eta_n+1=0 \tag{3.9}$$

(schéma implicite $\left(\begin{smallmatrix}\cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot\end{smallmatrix}\right)$), où

$$\mu_n^2=2\,\frac{\tau^2}{h^2}\sin^2\frac{n\pi h}{2}.$$

La résolution des équations du deuxième degré (3.8) et (3.9) donne respectivement

$$\eta_n=1-\mu_n^2\pm\sqrt{(1-\mu_n^2)^2-1}, \tag{3.10}$$

$$\eta_n=\frac{1}{1+\mu_n^2}\pm\sqrt{\left(\frac{1}{1+\mu_n^2}\right)^2-1}. \tag{3.11}$$

On voit facilement dans le cas du schéma de la croix que si

$$\mu_n^2 = 2\frac{\tau^2}{h^2}\sin^2\frac{n\pi h}{2} < 2, \tag{3.12}$$

alors

$$|\eta_{n_i}| = 1, \quad |\eta_{n_1} - \eta_{n_2}| > C\tau, \tag{3.13}$$

i.e. il y a stabilité. Il faut que la condition (3.12) ait lieu quel que soit $n = 1, 2, \ldots, N-1$. Cette condition se trouve évidemment remplie pour tout n si τ et h sont reliés par l'inégalité

$$\frac{\tau^2}{h^2} < \frac{1}{\sin^2\frac{\pi h n}{2}}, \tag{3.14}$$

ou par la condition plus simple de Courant

$$\frac{\tau}{h} < 1. \tag{3.15}$$

On établit sans peine que s'agissant du schéma implicite $\left(\begin{smallmatrix}\cdot & \cdot \\ \cdot & \cdot\end{smallmatrix}\right)$, on a l'égalité

$$|\eta_n| = 1 \tag{3.16}$$

quels que soient n et $\tau > 0$, i.e. le schéma est absolument stable (voir Richtmyer, Morton [III]).

Le problème de petites vibrations d'une membrane fait appel à des développements analogues. On a

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + f(x, y, t) \quad \text{dans} \quad D \times D_t \tag{3.17}$$

avec la condition aux limites

$$\varphi = g \quad \text{sur} \quad \partial D \times D_t \tag{3.18}$$

et les données initiales

$$\varphi = p, \qquad \frac{\partial\varphi}{\partial t} = q \quad \text{dans} \quad D \quad \text{pour} \quad t = 0.$$

On aborde ce problème dans le carré D pour des données initiales suffisamment régulières par une méthode des différences finies (on suit donc la même voie que dans le problème des cordes vibrantes) et on effectue une analyse moyennant les séries de Fourier comme pour l'analogue discrétisé de l'équation de la chaleur à deux variables. Nous en laissons le soin au lecteur et nous passons à un problème plus général pour les équations hyperboliques.

8.4. Equation du mouvement

Le transport de particules suivant une trajectoire, problème important en hydrodynamique, est gouverné par les équations du mouvement. Le fait que ces équations jouent un rôle absolument fondamental dans les applications explique leur attrait pour les chercheurs.

Les méthodes numériques associées sont développées selon les besoins de la dynamique des liquides et des gaz. Plusieurs idées intéressantes sont notamment dues à von Neumann, Richtmyer [VII], Dorodnitsyne [III; 5], Godounov [IV; 5], Lax [VI].

L'intérêt accru pour la prévision du temps et la dynamique de l'océan a stimulé l'élaboration de méthodes efficaces pour la résolution des équations du mouvement et on a formé, grâce aux résultats de Kurihara, Holloway [IV], Bryan [IV], Marchuk [III; 43a], des algorithmes universels performants pour ces problèmes.

L'un des problèmes principaux de la physique mathématique a trait à l'équation qui régit le mouvement suivant les trajectoires des particules:

$$\frac{d\varphi}{dt} = 0,$$

où

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z}$$

est la dérivée totale par rapport au temps de la fonction $\varphi(x, y, z, t)$; u, v et w sont les composantes du vecteur vitesse $u = ui + vj + wk$ telles que

$$u = \frac{dx}{dt}, \quad v = \frac{dy}{dt}, \quad w = \frac{dz}{dt}.$$

On résout l'équation proposée en introduisant des conditions supplémentaires dont la plus simple pour un milieu infini est

$$\varphi(x, y, z, 0) = f(x, y, z).$$

Un problème analogue intervient comme partie de l'algorithme général en hydrodynamique, en théorie du transfert de rayonnement et ailleurs, si bien qu'on va discuter en détail les tactiques numériques possibles.

8.4.1. Equation du mouvement en dimension 1

On rencontre souvent en dynamique et thermodynamique des fluides, dans les problèmes de météorologie et en dynamique de l'océan les équations qui gouvernent le transport de substance sui-

vant les trajectoires. Le cas le plus simple en est

$$\begin{gathered}\frac{\partial \varphi}{\partial t}+u \frac{\partial \varphi}{\partial x}=0 \quad \text{dans} \quad D \times D_t, \\ \varphi=f(x) \quad \text{dans} \quad D \quad \text{pour} \quad t=0,\end{gathered} \tag{4.1}$$

où u est la vitesse donnée et $f(x)$ la répartition initiale de φ. Le domaine D se confond avec l'axe réel des x et $D_t = \{0 \leqslant t \leqslant T\}$. On suppose que les fonctions $\varphi(x, t)$ et $f(x)$ sont périodiques de période 2π relativement à x. Si $u = \text{const}$, le problème proposé admet la solution évidente

$$\varphi(x, t) = f(x - ut) \tag{4.2}$$

à condition que $f(x)$ est dérivable. La solution (4.2) décrit la propagation de la perturbation initiale suivant les caractéristiques

$$x - ut = \text{const},$$

i.e. $\varphi(x, t) = \text{const}$ sur toute droite $x - ut = \text{const}$.

Ainsi, le problème (4.1) détermine pour $u > 0$ la propagation de la perturbation dans le sens des x croissants. On se rappelle ces résultats classiques lorsqu'on a à construire les analogues aux différences de (4.1). Si la vitesse $u = u(x, t)$ varie, la recherche analytique de la solution s'avère un problème délicat et on s'adresse à des méthodes numériques basées sur les approximations aux différences.

Considérons les schémas aux différences simples avec $u = \text{const}$. Pour fixer les idées, on pose $u > 0$, auquel cas on a le schéma explicite

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}+u \frac{\varphi_k^j-\varphi_{k-1}^j}{\Delta x}=0 \tag{4.3}$$

et le schéma implicite

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}+u \frac{\varphi_k^{j+1}-\varphi_{k-1}^{j+1}}{\Delta x}=0. \tag{4.4}$$

Les deux approchent à l'ordre 1 en Δx et τ. En effet, la valeur initiale de $f(x)$ et la solution $\varphi(x, t)$ seront supposées assez régulières. On développe la solution de (4.1) en série de Taylor autour du point $x = x_k$, $t = t_j$:

$$\varphi(x, t)=(\varphi)_k^j+(\varphi_t)_k^j(t-t_j)+(\varphi_x)_k^j(x-x_k)+\ldots \tag{4.5}$$

Portons dans (4.3), il vient

$$\frac{\partial \varphi}{\partial t}+u \frac{\partial \varphi}{\partial x}=\frac{u \Delta x}{2} \frac{\partial^2 \varphi}{\partial x^2}-\frac{\tau}{2} \frac{\partial^2 \varphi}{\partial t^2}. \tag{4.6}$$

Les termes supprimés sont d'ordre infinitésimal plus élevé. Il résulte de (4.1)

$$\frac{\partial^2 \varphi}{\partial t^2}=u^2 \frac{\partial^2 \varphi}{\partial x^2}, \tag{4.7}$$

auquel cas (4.6) se récrit

$$\frac{\partial \varphi}{\partial t} + u \frac{\partial \varphi}{\partial x} = \frac{u\Delta x - \tau u^2}{2} \frac{\partial^2 \varphi}{\partial x^2} \quad \text{pour} \quad x = x_k, \quad t = t_j. \tag{4.8}$$

Cette manière de procéder pour analyser les schémas aux différences est due à Joukov *). Voyons la relation (4.8) de plus près. On constate que lorsque

$$\frac{u\tau}{\Delta x} < 1,$$

elle s'interprète comme étant l'équation de la chaleur dont la solution est définie dans $x_{k-1} \leqslant x \leqslant x_k$, $t_j \leqslant t \leqslant t_{j+1}$. Sous l'hypothèse de petitesse des termes rejetés de (4.8), on obtient

$$\frac{\partial \varphi}{\partial t} + u \frac{\partial \varphi}{\partial x} = \mu \frac{\partial^2 \varphi}{\partial x^2},$$

où

$$\mu = \frac{u\Delta x - u^2\tau}{2}$$

est la *viscosité artificielle*. On note que si

$$\frac{u\tau}{\Delta x} = 1,$$

alors $\mu = 0$, ainsi que les autres termes rejetés; le schéma explicite (4.3) approche à tout ordre en Δx et τ.

Un cas intéressant est celui où

$$\frac{u\tau}{\Delta x} > 1,$$

ce qui détermine

$$\frac{\partial \varphi}{\partial t} + u \frac{\partial \varphi}{\partial x} = -|\mu| \frac{\partial^2 \varphi}{\partial x^2}. \tag{4.9}$$

Cette équation jointe à la condition initiale

$$\varphi = \varphi^0(x) \quad \text{pour} \quad t = 0 \tag{4.10}$$

conduit, on le vérifie sans peine, à un problème mal posé au sens d'Hadamard. La solution de ce problème est instable par rapport aux petites variations de données initiales. Lorsqu'on discrétise les équations des problèmes du type (4.1), on doit prendre en considération le critère de problème bien posé

$$\mu = \frac{u\tau}{\Delta x} \leqslant 1.$$

Etudions la propriété de stabilité numérique du schéma (4.3).

*) Voir Rojdestvenski et Yanenko [II; 7]. L'idée a été développée dans des articles de Yanenko et Chokin [VII; 8].

On considère le problème spectral

$$(A^h\omega)_k \equiv u\,\frac{\omega_k-\omega_{k-1}}{\Delta x}=\lambda\omega_k \tag{4.11}$$

sur l'intervalle discrétisé infini $D_h=(-\infty<x_k<\infty)$. La solution bornée dans D_h de (4.11) est de la forme

$$\omega_k=e^{ikp\Delta x}, \tag{4.12}$$

p étant un entier quelconque. On porte dans (4.11) et on obtient pour la valeur propre λ_p

$$\lambda_p=\frac{u}{\Delta x}\left(2\sin^2\frac{p\Delta x}{2}+i\sin p\Delta x\right). \tag{4.13}$$

L'équation (4.3) s'écrit sous forme opératorielle

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+A^h\varphi^j=0 \tag{4.14}$$

et la solution est cherchée comme étant

$$\varphi^j=\sum_{p=-\infty}^{\infty}\varphi_p^j e^{ikp\,\Delta x}, \tag{4.15}$$

où φ_p^j est le coefficient de Fourier de φ^j. On a pour les coefficients φ_p^j

$$\frac{\varphi_p^{j+1}-\varphi_p^j}{\tau}+\lambda_p\varphi_p^j=0.$$

D'où

$$\varphi_p^{j+1}=T_p\varphi_p^j,$$

$T_p=1-\tau\lambda_p$ étant l'opérateur de transition pour les coefficients de Fourier.

La condition pour que toutes les composantes de φ_p^j soient de module non croissant est

$$|1-\tau\lambda_p|\leqslant 1. \tag{4.16}$$

Cette inégalité a lieu si

$$\frac{u\tau}{\Delta x}\leqslant 1.$$

En effet,

$$|1-\tau\lambda_p|^2=\left(1-\frac{2u\tau}{\Delta x}\sin^2\frac{p\Delta x}{2}\right)^2+\left(\frac{u\tau}{\Delta x}\right)^2\sin^2 p\Delta x=$$

$$=1-4\frac{u\tau}{\Delta x}\sin^2\frac{p\Delta x}{2}+\left(\frac{u\tau}{\Delta x}\right)^2\left(4\sin^4\frac{p\Delta x}{2}+\sin^2 p\Delta x\right)=$$

$$=1-4\frac{u\tau}{\Delta x}\sin^2\frac{p\Delta x}{2}+4\left(\frac{u\tau}{\Delta x}\right)^2\sin^2\frac{p\Delta x}{2}\left(\sin^2\frac{p\Delta x}{2}+\cos^2\frac{p\Delta x}{2}\right)=$$

$$=1-4\sin^2\frac{p\Delta x}{2}\left(\frac{u\tau}{\Delta x}\right)\left(1-\frac{u\tau}{\Delta x}\right)\geqslant 0 \tag{4.17}$$

entraîne de suite la condition voulue.

Ainsi, on a établi une condition de stabilité numérique du schéma aux différences, qui coïncide dans notre cas avec la condition sous laquelle l'équation (4.8) est bien posée (on le vérifie aisément).

S'agissant du schéma implicite (4.4), on montre par les mêmes raisonnements qu'il approche à l'ordre 1 en Δx et τ. En développant en série de Taylor pour $x = x_k$ et $t = t_j$, on obtient l'« équation asymptotique »

$$\frac{\partial \varphi}{\partial t} + u \frac{\partial \varphi}{\partial x} = \mu \frac{\partial^2 \varphi}{\partial x^2}, \tag{4.18}$$

avec

$$\mu = \frac{u\Delta x + u^2\tau}{2}.$$

On peut déjà parler de la différence essentielle entre les relations (4.8) et (4.18). La dernière viscosité artificielle est toujours positive, si bien qu'avec des données initiales assez régulières, l'équation (4.18) est nécessairement bien posée. On montre aisément la stabilité de (4.4) pour tout rapport de pas (i.e. sa stabilité absolue). En effet, l'opérateur de transition de chaque coefficient de Fourier vaut

$$T_p = \frac{1}{1 + \tau\lambda_p}.$$

D'où

$$|T_p| \leqslant 1.$$

Voici deux autres schémas originaux d'emploi fréquent:

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} + u \frac{\varphi_{k+1}^{j+1} - \varphi_{k-1}^{j+1}}{2\Delta x} = 0; \tag{4.19}$$

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} + u \frac{\varphi_{k+1}^{j+1/2} - \varphi_{k-1}^{j+1/2}}{2\Delta x} = 0, \tag{4.20}$$

avec

$$\varphi_k^{j+1/2} = \frac{1}{2}(\varphi_k^{j+1} + \varphi_k^j).$$

L'approximation (4.19) est, on le voit facilement, d'ordre 1 en τ et d'ordre 2 en Δx. L'équation différentielle associée est de la forme (4.18), où

$$\mu = \frac{u^2\tau}{2}.$$

Quelle est la propriété de stabilité du schéma (4.19)? On considère le problème spectral

$$(A^h\omega)_k \equiv u \frac{\omega_{k+1} - \omega_{k-1}}{2\Delta x} = \lambda\omega_k \tag{4.21}$$

dont la solution est cherchée sous la forme (4.12). Il vient pour λ_p

$$\lambda_p = i \frac{u}{\Delta x} \sin p\Delta x. \tag{4.22}$$

Le schéma (4.19) entraîne de suite pour les coefficients de Fourier

$$\frac{\varphi_p^{j+1} - \varphi_p^j}{\tau} + \lambda_p \varphi_p^{j+1} = 0,$$

ou

$$\varphi_p^{j+1} = T_p \varphi^j, \tag{4.23}$$

avec

$$T_p = \frac{1}{1+\tau\lambda_p}$$

Compte tenu de (4.22),

$$T_p = \frac{1}{1 + i\frac{u\tau}{\Delta x} \sin p\Delta x},$$

donc

$$|T_p| = \frac{1}{\sqrt{1 + \left(\frac{u\tau}{\Delta x}\right)^2 \sin^2 p\,\Delta x}} \leqslant 1,$$

i.e. il y a stabilité absolue du schéma (4.19).

C'est le schéma de Crank-Nicholson (4.20) qui est particulièrement utile dans les applications. On établit sans difficulté qu'il approche à l'ordre 2 en Δx et τ et n'est pas dissipatif, i.e. $\mu = 0$ dans (4.18) et les termes rejetés sont en $\tau\,\Delta x$, τ^2 et Δx^2. Quant à la stabilité numérique, on a

$$T_p = \frac{1 - i\frac{u\tau}{2\Delta x} \sin p\,\Delta x}{1 + i\frac{u\tau}{2\Delta x} \sin p\,\Delta x}.$$

D'où

$$|T_p| = 1,$$

i.e. il y a stabilité absolue. Il y a intérêt à noter qu'en remplaçant dans (4.3) l'analogue discrétisé de $u\frac{\partial\varphi}{\partial x}$

$$u\frac{\varphi_k^j - \varphi_{k-1}^j}{\Delta x}$$

par

$$u\frac{\varphi_{k+1}^j - \varphi_k^j}{\Delta x},$$

le schéma obtenu s'avère instable pour $u > 0$ pour tout rapport de pas (démonstration immédiate).

Une méthode intéressante pour le problème (4.1) procède du schéma proposé par Landau, Meiman et Khalatnikov [IV; 14]

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}+\frac{u}{\Delta x}\left[\left(\frac{\varphi_k^{j+1}+\varphi_k^j}{2}\right)-\left(\frac{\varphi_{k-1}^{j+1}+\varphi_{k-1}^j}{2}\right)\right]=0. \quad (4.24)$$

Il s'agit d'une approximation d'ordre 2 en τ et d'ordre 1 en x définie par la relation de récurrence

$$\varphi_k^{j+1}=\frac{1-\frac{u\tau}{2\Delta x}}{1+\frac{u\tau}{2\Delta x}}\varphi_k^j+\frac{\frac{u\tau}{2\Delta x}}{1+\frac{u\tau}{2\Delta x}}(\varphi_{k-1}^{j+1}+\varphi_{k-1}^j). \quad (4.25)$$

On prouve par la méthode de Fourier la stabilité absolue au sens de von Neumann du schéma (4.24).

On construit un schéma analogue pour un problème de mouvement à plusieurs variables et on établit sa stabilité pour l'équation à coefficients constants.

Ci-dessus u a été supposé constant et positif. Si u est négatif, on a (4.1) par la substitution $x = -x$. Mais c'est le cas $u = u(x, t)$ qui offre le plus grand intérêt pour les applications. On montre par des raisonnements élémentaires la possibilité d'une instabilité même pour des schémas dissipatifs implicites. C'est surtout vrai des problèmes non linéaires. Le fait est qu'en développant la solution du problème aux différences et le coefficient $u(x_k, t_j)$ en série de Fourier et en faisant le produit, on obtient des composantes plus longues que les facteurs et d'autres qui sont plus courtes.

Il se peut que l'« énergie » quitte alors les régions des grandes longueurs d'onde pour les ondes de plus courte longueur, ce qui se solde par l'instabilité des calculs en dépit de la stabilité du schéma donné à coefficient constant. On dit de cette instabilité qu'elle est *non linéaire*. Elle intervient des fois dans les problèmes linéaires à coefficients variables. On conçoit donc que la construction de schémas aux différences associés aux équations non linéaires ou aux équations à coefficients variables insensibles à toute perturbation constitue un problème d'actualité. Dans la plupart des cas, on supprime l'instabilité numérique à l'aide de schémas dissipatifs associés à une viscosité artificielle μ déterminée. Malheureusement, ils approchent en général à l'ordre 1 en l'un des paramètres τ, Δx ou en les deux à la fois.

Un rôle de premier plan revient dans les applications à l'équation

$$\frac{\partial\varphi}{\partial t}+\frac{\partial u\varphi}{\partial x}=0, \quad (4.26)$$

où $u = u(x, t)$.

On associe à ces équations des schémas aux différences absolument stables qui approchent, pour certaines classes de coefficients, à l'ordre 1, voire 2. On les étudiera pour des équations du type (4.26) à plusieurs variables.

8.4.2. Equation du mouvement à coefficients variables en dimension 2

Soit, dans le plan (x, y), le problème de déterminer l'évolution d'un ensemble de particules suivant des trajectoires données. On aboutit, dans le cadre des méthodes de la mécanique des milieux continus, au problème

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} + v\frac{\partial \varphi}{\partial y} = 0 \quad \text{dans} \quad D \times D_t, \\ &\varphi(x, y, 0) = g \quad \text{dans} \quad D. \end{aligned} \tag{4.27}$$

Ici $u = u(x, y, t)$, $v = v(x, y, t)$.

Les composantes du vecteur vitesse u, v sont supposées vérifier à tout instant l'équation de continuité

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0. \tag{4.28}$$

Soit D le rectangle $\{0 \leqslant x \leqslant a,\ 0 \leqslant y \leqslant b\}$ et supposons que la solution du problème (4.27) à coefficients u, v est périodique et prend les mêmes valeurs sur les côtés opposés de D.

On met les équations d'évolution (4.27) sous forme opératorielle

$$\frac{\partial \varphi}{\partial t} + A\varphi = 0, \tag{4.29}$$

avec

$$\varphi = g \quad \text{pour} \quad t = 0, \quad A = u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y}.$$

On montre aisément que l'opérateur A est tel que $(A\varphi, \varphi) = 0$. En effet, on définit sur l'espace de Hilbert le produit scalaire

$$(A\varphi, \varphi) = \int_0^a dx \int_0^b dy \left(u\frac{\partial \varphi}{\partial x} + v\frac{\partial \varphi}{\partial y}\right)\varphi, \tag{4.30}$$

et l'expression sous $\int$ devient, compte tenu de (4.28),

$$\left(u\frac{\partial \varphi}{\partial x} + v\frac{\partial \varphi}{\partial y}\right)\varphi = \frac{\partial u\frac{\varphi^2}{2}}{\partial x} + \frac{\partial v\frac{\varphi^2}{2}}{\partial y}.$$

Donc

$$(A\varphi, \varphi) = \int_0^a dx \int_0^b dy \left(\frac{\partial u\frac{\varphi^2}{2}}{\partial x} + \frac{\partial v\frac{\varphi^2}{2}}{\partial y}\right), \tag{4.31}$$

d'où

$$(A\varphi,\ \varphi) = 0 \tag{4.32}$$

à partir de la condition de périodicité à la frontière de la solution. Ainsi, l'opérateur A permet des schémas aux différences absolument stables.

On décompose A de façon que chaque opérateur élémentaire A_α, $\alpha = 1, 2$, vérifie la condition

$$(A_\alpha\varphi,\ \varphi) = 0. \tag{4.33}$$

Dans ce cas, le schéma de splitting-up conduit à un schéma aux différences absolument stable exact à l'ordre 2.

La décomposition formelle de A en

$$A_1 = u\frac{\partial}{\partial x}, \qquad A_2 = v\frac{\partial}{\partial y} \tag{4.34}$$

ne respecte pas la condition (4.33). On vérifie facilement la validité de

$$(A_1\varphi,\ \varphi) = -\frac{1}{2}\int_0^a dx \int_0^b \varphi^2 \frac{\partial u}{\partial x}\,dy, \qquad (A_2\varphi,\ \varphi) = -\frac{1}{2}\int_0^a dx \int_0^b \varphi^2 \frac{\partial v}{\partial y}\,dy,$$

i.e. avec ces A_1 et A_2 on ne construit pas de schéma de splitting-up successif.

Si l'on prend A_1, A_2 plus compliqués:

$$A_1\varphi = u\frac{\partial\varphi}{\partial x} + \frac{\varphi}{2}\frac{\partial u}{\partial x}, \qquad A_2\varphi = v\frac{\partial\varphi}{\partial y} + \frac{\varphi}{2}\frac{\partial v}{\partial y}, \tag{4.35}$$

on constate qu'on est dans la condition (4.33) et que leur somme vaut exactement A. En effet, on a

$$(A_1 + A_2)\,\varphi = u\frac{\partial\varphi}{\partial x} + v\frac{\partial\varphi}{\partial y} + \frac{\varphi}{2}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = u\frac{\partial\varphi}{\partial x} + v\frac{\partial\varphi}{\partial y} = A\varphi.$$

On a tenu compte de ce que les coefficients u et v satisfont à (4.28).

Ainsi, toutes les conditions nécessaires d'application de la méthode considérée se trouvent vérifiées et on est conduit au schéma de décomposition sur l'intervalle $t_{j-1} \leqslant t \leqslant t_{j+1}$:

$$\begin{aligned}
&\frac{\varphi^{j-1/2} - \varphi^{j-1}}{\tau} + \left(u^j\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial u^j}{\partial x}\right)\frac{\varphi^{j-1/2} + \varphi^{j-1}}{2} = 0,\\
&\frac{\varphi^{j} - \varphi^{j-1/2}}{\tau} + \left(v^j\frac{\partial}{\partial y} + \frac{1}{2}\frac{\partial v^j}{\partial y}\right)\frac{\varphi^{j} + \varphi^{j-1/2}}{2} = 0,\\
&\frac{\varphi^{j+1/2} - \varphi^{j}}{\tau} + \left(v^j\frac{\partial}{\partial y} + \frac{1}{2}\frac{\partial v^j}{\partial y}\right)\frac{\varphi^{j+1/2} + \varphi^{j}}{2} = 0,\\
&\frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + \left(u^j\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial u^j}{\partial x}\right)\frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} = 0.
\end{aligned} \tag{4.36}$$

En cas de u, v et de la solution φ suffisamment réguliers par rapport à toutes les variables, le schéma (4.36) approche le problème initial à l'ordre 2 et présente la propriété de stabilité absolue au sens de

$$\| \varphi^{j+1} \| = \| \varphi^{j-1} \| = \ldots = \| g \|. \tag{4.37}$$

L'exemple considéré montre on ne peut plus clairement qu'une décomposition formelle (4.34) risque de compromettre l'idée même de la méthode, si bien que des schémas justifiés et efficaces n'apparaissent que par un jeu de considérations complémentaires.

Après cette sorte de préambule on passe à la discrétisation de (4.27) par rapport aux variables spatiales et au temps. Comment approcher au mieux l'opérateur A par rapport à x et y? On connaît dès le chapitre 2 l'intérêt de l'approximation séquentielle qui conserve la qualité et les propriétés d'additivité des opérateurs.

Supposons les coefficients u et v suffisamment réguliers et mettons les équations (4.27) sous forme conservative

$$\begin{gathered} \frac{\partial \varphi}{\partial t} + \frac{\partial u \varphi}{\partial x} + \frac{\partial v \varphi}{\partial y} = 0 \quad \text{dans} \quad D \times D_t, \\ \varphi = g \quad \text{dans} \quad D \quad \text{pour} \quad t = 0. \end{gathered} \tag{4.38}$$

On construit le schéma aux différences à partir de l'opérateur A défini par

$$A\varphi = \frac{\partial u \varphi}{\partial x} + \frac{\partial v \varphi}{\partial y} - \frac{\varphi}{2} \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right). \tag{4.39}$$

Son analogue discrétisé s'écrit

$$(A^h \varphi)_{k,\, l} = \frac{u_{k+1,\, l} \varphi_{k+1,\, l} - u_{k-1,\, l} \varphi_{k-1,\, l}}{2\Delta x} + \frac{v_{k,\, l+1} \varphi_{k,\, l+1} - v_{k,\, l-1} \varphi_{k,\, l-1}}{2\Delta y} - \\ - \frac{\varphi_{k,\, l}}{2} \left(\frac{u_{k+1,\, l} - u_{k-1,\, l}}{2\Delta x} + \frac{v_{k,\, l+1} - v_{k,\, l-1}}{2\Delta y} \right). \tag{4.40}$$

Evidemment, cette relation approche (4.39) à l'ordre 2 en Δx et Δy pour u, v et φ suffisamment réguliers. L'inconvénient majeur en est que l'opérateur A^h ainsi défini n'est plus symétrique gauche, i.e.

$$(A^h \varphi,\ \varphi) \neq 0. \tag{4.41}$$

Ainsi, l'approximation en question tombe en défaut.

On montre que l'approximation suivante de (4.39)

$$(A^h \varphi)_{k,\, l} = \frac{u_{k+1/2,\, l} \varphi_{k+1,\, l} - u_{k-1/2,\, l} \varphi_{k-1,\, l}}{2\Delta x} + \\ + \frac{v_{k,\, l+1/2} \varphi_{k,\, l+1} - v_{k,\, l-1/2} \varphi_{k,\, l-1}}{2\Delta y} \tag{4.42}$$

respecte la relation fondamentale

$$(A^h \varphi,\ \varphi) = 0 \tag{4.43}$$

et est d'ordre 2 en Δx et Δy. On a pour les coefficients

$$\begin{aligned} u_{k+1/2,\,l} &= u_{k+1,\,l} - \frac{u_{k+1,\,l} - u_{k,\,l}}{2}\,; \\ v_{k,\,l+1/2} &= v_{k,\,l+1} - \frac{v_{k,\,l+1} - v_{k,\,l}}{2}\,; \\ u_{k-1/2,\,l} &= u_{k-1,\,l} + \frac{u_{k,\,l} - u_{k-1,\,l}}{2}\,; \\ v_{k,\,l-1/2} &= v_{k,\,l-1} + \frac{v_{k,\,l} - v_{k,\,l-1}}{2}\,. \end{aligned} \tag{4.44}$$

On porte ces expressions dans (4.42) et on obtient par de simples transformations

$$\begin{aligned} (A^h\varphi)_{k,\,l} = {} & \frac{u_{k+1,\,l}\varphi_{k+1,\,l} - u_{k-1,\,l}\varphi_{k-1,\,l}}{2\Delta x} + \frac{v_{k,\,l+1}\varphi_{k,\,l+1} - v_{k,\,l-1}\varphi_{k,\,l-1}}{2\Delta y} - \\ & - \frac{\varphi_{k,\,l}}{2}\left(\frac{u_{k+1,\,l} - u_{k-1,\,l}}{2\Delta x} + \frac{v_{k,\,l+1} - v_{k,\,l-1}}{2\Delta y}\right) - \\ & - (\Delta x^2 R_{k,\,l} + \Delta y^2 Q_{k,\,l}), \end{aligned} \tag{4.45}$$

avec

$$R_{k,\,l} \to \frac{1}{4}\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\frac{\partial \varphi}{\partial x}\right), \qquad Q_{k,\,l} \to \frac{1}{4}\frac{\partial}{\partial y}\left(\frac{\partial v}{\partial y}\frac{\partial \varphi}{\partial y}\right)$$

lorsque $\Delta x \to 0$ et $\Delta y \to 0$.

On suppose maintenant $u_{k,l}$ et $v_{k,l}$ tels que

$$\frac{u_{k+1,\,l} - u_{k-1,\,l}}{2\Delta x} + \frac{v_{k,\,l+1} - v_{k,\,l-1}}{2\Delta y} = O(h^2). \tag{4.46}$$

Si les coefficients u, v et la solution φ admettent des dérivées troisièmes par rapport à x et à y bornées, l'écart entre (4.45) avec (4.46) et (4.40) est du même ordre infinitésimal que celui entre (4.40) et (4.39). Ainsi, (4.42) approche (4.39) à l'ordre 2 en Δx et Δy.

On montre que A^h vérifie la condition (4.43) et qu'on a

$$(A^h_\alpha \varphi,\ \varphi) = 0 \tag{4.47}$$

pour A_1^h et A_2^h définis par les relations

$$\begin{aligned} A_1^h\varphi &= \frac{u_{k+1/2,\,l}\varphi_{k+1,\,l} - u_{k-1/2,\,l}\varphi_{k-1,\,l}}{2\Delta x}\,, \\ A_2^h\varphi &= \frac{v_{k,\,l+1/2}\varphi_{k,\,l+1} - v_{k,\,l-1/2}\varphi_{k,\,l-1}}{2\Delta y}\,. \end{aligned} \tag{4.48}$$

On introduit le produit scalaire des vecteurs a et b

$$(a,\ b) = \sum_k \sum_l a_{k,\,l} b_{k,\,l} \Delta x \Delta y,$$

donc

$$(A_1^h\varphi,\ \varphi) =$$
$$= \frac{1}{2}\sum_k\sum_l \Delta y\left(\sum_k u_{k+1/2,\,l}\varphi_{k+1,\,l} - \sum_k u_{k-1/2,\,l}\varphi_{k-1,\,l}\right)\varphi_{k,\,l},$$
$$(A_2^h\varphi,\ \varphi) = \frac{1}{2}\sum_l\sum_k \Delta x\left(\sum_l v_{k,\,l+1/2}\varphi_{k,\,l+1} - \sum_l v_{k,\,l-1/2}\varphi_{k,\,l-1}\right)\varphi_{k,\,l}. \tag{4.49}$$

Réduisons les termes semblables, il vient (4.47). Cette condition entraîne de suite (4.43). Ainsi, on a réalisé l'approximation nécessaire par rapport aux variables d'espace. Il reste à discrétiser dans le temps le système d'équations différentielles ordinaires

$$\frac{d\varphi^h}{dt} + A^h\varphi^h = 0 \quad \text{dans} \quad D_h \times D_t,$$
$$\varphi^h = g^h \quad \text{dans} \quad D_h \quad \text{pour} \quad t = 0, \tag{4.50}$$

avec

$$A^h = A_1^h + A_2^h,$$

φ^h une fonction vectorielle de composantes $\varphi_{k,l}$ et A_α^h tel qu'on ait (4.47). Cela signifie que le problème (4.50) se prête à la décomposition. L'indice h des fonctions et des opérateurs étant sans importance est omis, et on aboutit sur l'intervalle $t_{j-1} \leqslant t \leqslant t_{j+1}$ au système

$$\frac{\varphi^{j-1/2} - \varphi^{j-1}}{\tau} + A_1^j\frac{\varphi^{j-1/2} + \varphi^{j-1}}{2} = 0,$$
$$\frac{\varphi^{j} - \varphi^{j-1/2}}{\tau} + A_2^j\frac{\varphi^{j} + \varphi^{j-1/2}}{2} = 0,$$
$$\frac{\varphi^{j+1/2} - \varphi^{j}}{\tau} + A_2^j\frac{\varphi^{j+1/2} + \varphi^{j}}{2} = 0,$$
$$\frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + A_1^j\frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} = 0. \tag{4.51}$$

Ainsi, on a ramené le problème (4.27) à un système d'équations aux différences unidimensionnelles simples dont la solution est cherchée par factorisation des équations discrétisées à trois points.

Les expériences numériques correspondantes sont données dans l'article de Marchouk, Rivine et Youdine [IV; 18].

8.4.3. Equation du mouvement à plusieurs variables

Soit l'équation qui gouverne l'évolution des fluides suivant une trajectoire

$$\frac{\partial\Phi}{\partial t} + \sum_{\alpha=1}^{n} v_\alpha\frac{\partial\Phi}{\partial x_\alpha} = 0 \quad \text{dans} \quad D \times D_t,$$
$$\Phi(x,\ 0) = f(x) \quad \text{dans} \quad D. \tag{4.52}$$

Les coefficients sont supposés vérifier l'équation de continuité

$$\sum_{\alpha=1}^{n} \frac{\partial v_\alpha}{\partial x_\alpha} = 0 \quad \text{dans} \quad D \times D_t. \tag{4.53}$$

Le problème (4.52) avec la condition (4.53) est mis sous forme conservative

$$\begin{gathered} \frac{\partial \Phi}{\partial t} + \sum_{\alpha=1}^{n} \frac{\partial v_\alpha \Phi}{\partial x_\alpha} = 0 \quad \text{dans} \quad D \times D_t, \\ \Phi(x, 0) = f(x) \quad \text{dans} \quad D. \end{gathered} \tag{4.54}$$

On discrétise l'équation (4.54) moyennant le schéma de Crank-Nicholson

$$\frac{\Phi^{j+1} - \Phi^j}{\tau} + \sum_{\alpha=1}^{n} \frac{\partial u_\alpha^j \Phi^{j+1/2}}{\partial x_\alpha} = 0 \qquad (t_j \leqslant t \leqslant t_{j+1}), \tag{4.55}$$

où

$$\Phi^{j+1/2} = \frac{1}{2}(\Phi^{j+1} + \Phi^j). \tag{4.56}$$

Ce schéma a utilisé une approximation des coefficients v_α. Cette approximation doit être naturellement d'ordre 1 ou 2 en τ. Dans le premier cas, on prend par exemple

$$u_\alpha^j = v_\alpha(x, t_j),$$

et dans le second

$$u_\alpha^j = \frac{v_\alpha(x, t_{j+1}) + v_\alpha(x, t_j)}{2}.$$

On introduit la notation

$$A^j \Phi = \sum_{\alpha=1}^{n} A_\alpha^j \Phi, \qquad A_\alpha^j \Phi = \frac{\partial u_\alpha^j \Phi}{\partial x_\alpha},$$

et (4.55) se récrit

$$\frac{\Phi^{j+1} - \Phi^j}{\tau} + A^j \Phi^{j+1/2} = 0, \qquad \Phi^{j+1/2} = \frac{1}{2}(\Phi^{j+1} + \Phi^j)$$

ou

$$\left(E + \frac{\tau}{2} A^j\right) \Phi^{j+1} = \left(E - \frac{\tau}{2} A^j\right) \Phi^j. \tag{4.57}$$

On suppose, pour simplifier l'exposé, que la solution du problème est périodique relativement au parallélépipède D de dimension n.

On introduit le produit scalaire

$$(a,\ b) = \int_D ab\, dD$$

et on établit sans peine la validité de l'égalité

$$(A^j\Phi,\ \Phi) = 0. \tag{4.58}$$

On résout l'équation (4.57) par rapport à Φ^{j+1}:

$$\Phi^{j+1} = \left(E + \frac{\tau}{2} A^j\right)^{-1} \left(E - \frac{\tau}{2} A^j\right) \Phi^j,$$

et on a à l'aide du lemme du Kellogg et de la condition (4.58)

$$\| \Phi^{j+1} \| = \| \Phi^j \|. \tag{4.59}$$

Considérons l'approximation aux différences de (4.55) dans l'espace des variables géométriques. On projette D sur D_h et on écrit en omettant les indices constants de u, Φ:

$$\frac{\Phi^{j+1} - \Phi^j}{\tau} + \Lambda^j \Phi^{j+1/2} = 0, \tag{4.60}$$

où

$$\Lambda^j = \sum_{\alpha=1}^{n} \Lambda_\alpha^j, \tag{4.61}$$

$$\Lambda_\alpha^j \Phi = \frac{1}{2\Delta x_\alpha} (u^j_{\alpha,\, k_\alpha+1/2} \Phi_{k_\alpha+1} - u^j_{\alpha,\, k_\alpha-1/2} \Phi_{k_\alpha-1}), \tag{4.62}$$

l'indice k_α correspondant aux numéros des nœuds associés à la variable x_α. On introduit le produit scalaire

$$(a,\ b) = \sum_{k_1 \ldots k_n} a_{k_1 k_2 \ldots k_n} b_{k_1 k_2 \ldots k_n} \Delta x_1 \Delta x_2 \ldots \Delta x_n \tag{4.63}$$

et on vérifie aisément que

$$(\Lambda^j\Phi,\ \Phi) = 0. \tag{4.64}$$

On a donc, compte tenu de cette condition:

$$\| \Phi^{j+1} \| = \| \Phi^j \| = \ldots = \| f \|,$$

Φ^j étant solution de (4.60). Ainsi, l'approximation indiquée de A^j par Λ^j de (4.60)-(4.62) conduit encore à un schéma absolument stable.

Etudions cette approximation. On considère à cet effet les opérateurs élémentaires A_α^j et Λ_α^j et on prend pour coefficients de l'expression de $\Lambda_\alpha^j \Phi$

$$\begin{aligned} u^j_{\alpha,\, k_\alpha+1/2} &= u^j_{\alpha,\, k_\alpha+1} - \frac{1}{2}(u^j_{\alpha,\, k_\alpha+1} - u^j_{\alpha,\, k_\alpha}), \\ u^j_{\alpha,\, k_\alpha-1/2} &= u^j_{\alpha,\, k_\alpha-1} + \frac{1}{2}(u^j_{\alpha,\, k_\alpha} - u^j_{\alpha,\, k_\alpha-1}), \end{aligned} \tag{4.65}$$

ce qui donne

$$\Lambda_\alpha^j \Phi = \frac{u^j_{\alpha,\, k_\alpha+1}\Phi_{\alpha,\, k_\alpha+1} - u^j_{\alpha,\, k_\alpha-1}\Phi_{k_\alpha-1}}{2\Delta x_\alpha} - \\ - \frac{\Phi_{k_\alpha}}{2}\,\frac{u^j_{\alpha,\, k_\alpha+1} - u^j_{\alpha,\, k_\alpha-1}}{2\Delta x_\alpha} - \frac{\Delta x_\alpha^2}{4} R_\alpha^j, \qquad (4.66)$$

où

$$R_\alpha^j = \frac{1}{\Delta x_\alpha^3}\left[(u^j_{\alpha,\, k_\alpha+1} - u^j_{\alpha,\, k_\alpha})\,(\Phi_{k_\alpha+1} - \Phi_{k_\alpha}) - \right. \\ \left. - (u^j_{\alpha,\, k_\alpha} - u^j_{\alpha,\, k_\alpha-1})\,(\Phi_{k_\alpha} - \Phi_{k_\alpha-1})\right]. \qquad (4.67)$$

Etant donnée la régularité suffisante de la solution et des coefficients $u\,(x,\, t)$, si $\Delta x_\alpha \to 0$, alors

$$R_\alpha^j \to \frac{\partial}{\partial x_\alpha}\left(\frac{\partial u^j}{\partial x_\alpha}\,\frac{\partial \Phi}{\partial x_\alpha}\right).$$

Il résulte de (4.66) que Λ_α^j n'approche en général pas A_α^j.

Considérons Λ^j et l'expression de $\Lambda^j\Phi$. On a

$$\Lambda^j\Phi = \sum_{\alpha=1}^{n} \frac{u^j_{\alpha,\, k_\alpha+1}\Phi_{k_\alpha+1} - u^j_{\alpha,\, k_\alpha-1}\Phi_{k_\alpha-1}}{2\Delta x_\alpha} - \\ - \sum_{\alpha=1}^{n} \frac{\Phi_{k\alpha}}{2}\,\frac{u^j_{\alpha,\, k_\alpha+1} - u^j_{\alpha,\, k_\alpha-1}}{2\Delta x_\alpha} - \frac{1}{4}\sum_{\alpha=1}^{n} \Delta x_\alpha^2 R_\alpha^j. \qquad (4.68)$$

Supposons l'équation de continuité discrétisée de façon que

$$\sum_{\alpha=1}^{n} \frac{u_{\alpha,\, k_\alpha+1} - u_{\alpha,\, k_\alpha-1}}{2\Delta x_\alpha} = O\,(h^2), \qquad (4.69)$$

où $h = \max\limits_{\alpha} \{\Delta x_\alpha\}$. La deuxième somme de (4.68) s'annule, si bien que

$$\Lambda^j\Phi = \sum_{\alpha=1}^{n} \frac{u^j_{\alpha,\, k_\alpha+1}\Phi_{k_\alpha+1} - u^j_{\alpha,\, k_\alpha-1}\Phi_{k_\alpha-1}}{2\Delta x_\alpha} + O\,(h^2), \qquad (4.70)$$

i.e. Λ^j approche A à des infiniment petits du second ordre près en toutes les variables géométriques.

On passe à la décomposition du problème (4.54). Soit le schéma à 2 cycles

$$\begin{gathered}\left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j-\frac{n-\alpha}{n}}=\left(E-\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j-\frac{n-\alpha+1}{n}},\\ \alpha=1,\ 2,\ \ldots,\ n,\\ \left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+\frac{n-\alpha+1}{n}}=\left(E-\frac{\tau}{2}\Lambda_\alpha^j\right)\Phi^{j+\frac{n-\alpha}{n}},\\ \alpha=n,\ n-1,\ \ldots,\ 1.\end{gathered}\tag{4.71}$$

Il faut que les coefficients de Λ_α^j soient $u^j = v\,(x,\ t_j)$, ce qui garantit une approximation d'ordre 2 sur chaque intervalle

$$t_{j-1}\leqslant t\leqslant t_{j+1}.\tag{4.72}$$

La méthode (4.71) est absolument stable. En effet,

$$(\Lambda_\alpha^j\Phi,\ \Phi)=0,\tag{4.73}$$

si bien qu'on démontre facilement par le lemme de Kellogg que

$$\|T_\alpha^j\|=1,\qquad \alpha=1,\ 2,\ \ldots,\ n,$$

et donc

$$\begin{aligned}&\left\|\Phi^{j-\frac{n-\alpha}{n}}\right\|=\left\|\Phi^{j-\frac{n-\alpha+1}{n}}\right\|,\qquad \alpha=1,\ 2,\ \ldots,\ n,\\ &\left\|\Phi^{j+\frac{n-\alpha+1}{n}}\right\|=\left\|\Phi^{j+\frac{n-\alpha}{n}}\right\|,\qquad \alpha=n,\ n-1,\ \ldots,\ 1.\end{aligned}$$

On a par élimination des valeurs intermédiaires de $\|\Phi\|$

$$\|\Phi^{j+1}\|=\|\Phi^{j-1}\|.\tag{4.74}$$

La méthode considérée pour les problèmes de mouvement en mécanique des fluides se transpose aisément aux équations quasi linéaires à condition d'adjoindre au schéma de résolution un bon schéma d'extrapolation sur les coefficients u_α au temps t_j connaissant leurs valeurs aux époques antérieures.

L'algorithme joue pour des approximations plus bonnes que des schémas d'ordre 2 en les variables d'espace. Soit, en effet,

$$\Lambda_\alpha\Phi=\sum_{m=1}^{p}\beta_m\frac{\frac{1}{2}(u_{k+m}+u_k)\Phi_{k+m}-\frac{1}{2}(u_k+u_{k-m})\Phi_{k-m}}{2(m\Delta x_\alpha)},\tag{4.75}$$

avec β_m vérifiant le système d'équations

$$\sum_{m=1}^{p}\beta_m=1,\quad \sum_{m=1}^{p}m^2\beta_m=0,\ \ldots,\ \sum_{m=1}^{p}m^{2p-2}\beta_m=0.\tag{4.76}$$

Dans ce cas,

$$\sum_{\alpha=1}^{n} \Lambda_\alpha \Phi = \Big(\sum_{\alpha=1}^{n} \frac{\partial u_\alpha \Phi}{\partial x_\alpha} \Big)_{x_\alpha = x_{k_\alpha}} + O(h^{2p})$$

si $x_\alpha = x_{k_\alpha}$, et on suppose que les coefficients satisfont à la relation

$$\sum_{\alpha=1}^{n} \sum_{m=1}^{p} \beta_m \frac{u_{\alpha, k_\alpha+m} - u_{\alpha, k_\alpha-m}}{2m\Delta x_\alpha} = O(h^{2p}). \tag{4.77}$$

On utilise l'algorithme de décomposition (4.71) et on obtient une solution en $O(h^{2p} + \tau^2)$ du problème (4.54).

On applique le même algorithme s'il y a compressibilité du milieu, auquel cas le système fondamental d'équations s'écrit

$$\begin{gathered} \frac{\partial \rho \Phi}{\partial t} + \sum_{\alpha=1}^{n} \frac{\partial \rho u_\alpha \Phi}{\partial x_\alpha} = 0, \\ \frac{\partial \rho}{\partial t} + \sum_{\alpha=1}^{n} \frac{\partial \rho u_\alpha}{\partial x_\alpha} = 0. \end{gathered} \tag{4.78}$$

Comme ρ est en fait une fonction positive, le système se ramène à

$$\begin{gathered} \frac{\partial \sqrt{\rho}\, \Phi}{\partial t} + \sum_{\alpha=1}^{n} \Big(\frac{1}{2} u_\alpha \frac{\partial \sqrt{\rho}\, \Phi}{\partial x_\alpha} + \frac{1}{2} \frac{\partial \sqrt{\rho}\, u_\alpha \Phi}{\partial x_\alpha} \Big) = 0, \\ \frac{\partial \sqrt{\rho}}{\partial t} + \sum_{\alpha=1}^{n} \Big(\frac{1}{2} u_\alpha \frac{\partial \sqrt{\rho}}{\partial x_\alpha} + \frac{1}{2} \frac{\partial \sqrt{\rho}\, u_\alpha}{\partial x_\alpha} \Big) = 0 \end{gathered} \tag{4.79}$$

(Siniaev [IV; 26], Marchouk, Dymnikov et al. [IV; 17], Pénenko [IV; 21]). On note $\sqrt{\rho}\Phi = \psi$ et on approche (4.79) par

$$\begin{gathered} \frac{\psi^{j+1} - \psi^j}{\tau} + \sum_{\alpha=1}^{n} \frac{(u_\alpha)_{k_\alpha+1/2} \psi^{j+1/2}_{k_\alpha+1} - (u_\alpha)_{k_\alpha-1/2} \psi^{j+1/2}_{k_\alpha-1}}{2\Delta x_\alpha} = 0, \\ \frac{\sqrt{\rho}^{j+1} - \sqrt{\rho}^{j}}{\tau} + \sum_{\alpha=1}^{n} \frac{(u_\alpha)_{k_\alpha+1/2} \sqrt{\rho}^{j+1/2}_{k_\alpha+1} - (u_\alpha)_{k_\alpha-1/2} \sqrt{\rho}^{j+1/2}_{k_\alpha-1}}{2\Delta x_\alpha} = 0, \end{gathered} \tag{4.80}$$

où

$$(u_\alpha)_{k_\alpha+1/2} = \frac{(u_\alpha)_{k_\alpha+1} + (u_\alpha)_{k_\alpha-1}}{2},$$

$$\psi^{j+1} = \frac{\psi^{j+1/2} + \psi^j}{2}.$$

L'approximation est, on le voit facilement, d'ordre 2 en τ et h, et on a les relations

$$\| \psi^{j+1} \| = \| \psi^j \|, \qquad \| (\sqrt{\rho}^{j+1}) \| = \| (\sqrt{\rho}^j) \|.$$

On aborde le problème (4.80) par une technique analogue à (4.71).

8.5. Equation de transport non stationnaire

Dans ce numéro, on applique l'idée de décomposition à la théorie du transfert de rayonnement, secteur de pointe de la physique mathématique.

Comme notre but numéro un est d'étudier les méthodes, nous nous bornerons à plusieurs modèles simples mais d'un grand intérêt pratique. Bien que l'appareil mathématique de la résolution repose essentiellement sur l'idée de décomposition, les principes de base de la discrétisation des problèmes de transport se conservent pour d'autres méthodes.

Soit le problème de transport en géométrie plane avec les données initiales

$$\frac{1}{c}\frac{\partial \varphi}{\partial t} + \mu \frac{\partial \varphi}{\partial z} + \sigma \varphi = \frac{\sigma_s}{2} \int_{-1}^{1} \varphi \, d\mu + f, \tag{5.1}$$

$$\varphi = 0 \quad \text{pour} \quad z = 0, \ \mu > 0, \tag{5.2}$$

$$\varphi = 0 \quad \text{pour} \quad z = H, \ \mu < 0,$$

$$\varphi = \varphi^0 \quad \text{pour} \quad t = 0, \tag{5.3}$$

où $\varphi(z, \mu, t)$ est la densité en z à l'instant t des particules volant à la vitesse c *) suivant une trajectoire qui fait l'angle ν avec l'axe oz; $\mu = \cos \nu$; $f(z, \mu, t)$ sont des sources de rayonnement données et $\sigma(z)$, $\sigma_s(r)$ des fonctions continues par morceaux telles que

$$0 < \sigma_0 \leqslant \sigma \leqslant \sigma_1 < \infty, \qquad 0 \leqslant \sigma_s \leqslant \sigma_s' < \infty,$$

$$0 < \sigma_{c0} \leqslant \sigma_c = \sigma_c - \sigma_s.$$

Effectuons les transformations introduites par Kouznétsov (voir Kouznétsov, Marchouk [XVII; 8]). On doit à Vladimirov [XVII; 3] des transformations analogues pour le problème à plusieurs variables. Désignons par φ^+ la solution de (5.1) pour $\mu > 0$ et par φ^- dans le cas contraire. L'équation de transport se met sous forme de système

*) Dans le souci de simplifier le calcul, on pose $c \equiv 1$.

de deux équations

$$\begin{aligned}\frac{\partial\varphi^+}{\partial t}+\mu\frac{\partial\varphi^+}{\partial z}+\sigma\varphi^+&=\frac{\sigma_s}{2}\int_0^1(\varphi^++\varphi^-)\,d\mu'+f^+,\\ \frac{\partial\varphi^-}{\partial t}-\mu\frac{\partial\varphi^-}{\partial z}+\sigma\varphi^-&=\frac{\sigma_s}{2}\int_0^1(\varphi^++\varphi^-)\,d\mu'+f^-.\end{aligned}\tag{5.4}$$

φ^+ et φ^- satisfont aux conditions aux limites respectives

$$\begin{aligned}\varphi^+(z,\ \mu)&=0 \quad \text{pour} \quad z=0,\\ \varphi^-(z,\ \mu)&=0 \quad \text{pour} \quad z=H.\end{aligned}\tag{5.5}$$

On additionne les deux équations et on fait leur différence. Il vient

$$\frac{\partial u}{\partial t}+\mu\frac{\partial v}{\partial z}+\sigma u=\sigma_s\int_0^1 u\,d\mu'+g,\tag{5.6}$$

$$\frac{\partial v}{\partial t}+\mu\frac{\partial u}{\partial z}+\sigma v=r,$$

où

$$u=\frac{1}{2}(\varphi^++\varphi^-),\qquad v=\frac{1}{2}(\varphi^+-\varphi^-),$$

$$g=\frac{1}{2}(f^++f^-),\qquad r=\frac{1}{2}(f^+-f^-).$$

On vérifie aisément que les conditions aux limites (5.5) deviennent

$$\begin{aligned}u+v&=0 \quad \text{pour} \quad z=0,\\ u-v&=0 \quad \text{pour} \quad z=H,\end{aligned}\tag{5.7}$$

et les données initiales sont

$$u=u^0,\quad v=v^0 \quad \text{pour} \quad t=0.\tag{5.8}$$

On met le problème (5.6)-(5.8) sous forme opératorielle. On introduit les fonctions vectorielles w, w^0, F et l'opérateur A définis comme suit :

$$w=\left\|\begin{matrix}u\\ v\end{matrix}\right\|,\quad w^0=\left\|\begin{matrix}u^0\\ v^0\end{matrix}\right\|,\quad F=\left\|\begin{matrix}g\\ r\end{matrix}\right\|,$$

$$A=\left\|\begin{matrix}\sigma-\sigma_s\displaystyle\int_0^1 d\mu' & \mu\dfrac{\partial}{\partial z}\\ \mu\dfrac{\partial}{\partial z} & \sigma\end{matrix}\right\|\tag{5.9}$$

et, dans $D \equiv [0, H] \times [0, 1]$, l'espace de Hilbert de fonctions $L_2(D)$ à produit scalaire

$$(a, b) = \sum_{i=1}^{2} \int_0^1 d\mu \int_0^H a^i b^i \, dx, \tag{5.10}$$

a^i, b^i étant les composantes des fonctions vectorielles a et b.

On définit dans cet espace le sous-espace Φ de fonctions vectorielles dont les éléments vérifient la condition

$$(Aw, w) < +\infty. \tag{5.11}$$

Ce produit scalaire sera, sauf indication contraire, supposé dépendre du temps. On impose aux composantes des fonctions vectorielles w d'être continues dans D et d'admettre dans ce domaine des dérivées premières $\frac{\partial w}{\partial z}$ absolument continues. On note que la régularité dans D de u et v découle de suite de la condition de régularité dans D des fonctions φ^+ et φ^-. On définit enfin dans Φ, et on note Φ^0 un ensemble de fonctions vectorielles qui vérifient (5.7) et possèdent une dérivée première par rapport au temps absolument continue. Evidemment Φ^0 est le domaine de l'opérateur

$$L \equiv \frac{\partial}{\partial t} + A.$$

On est donc conduit au problème

$$\begin{aligned} &\frac{\partial w}{\partial t} + Aw = F \quad \text{dans} \quad D \times [0, T], \\ &w = w^0 \quad \text{pour} \quad t = 0 \quad \text{dans} \quad D, \end{aligned} \tag{5.12}$$

et

$$F(t) \in L_2(D \times [0, T]), \quad w^0 \in \Phi, \; w(t) \in \Phi^0.$$

On vérifie sans peine qu'on a pour les fonctions de Φ^0 domaine de A la relation

$$(Aw, w) > 0. \tag{5.13}$$

Vladimirov a de plus montré (voir [XVII ; 3]) la définie positivité de A, i.e.

$$(Aw, w) \geqslant \gamma (w, w), \tag{5.14}$$

γ étant une constante positive liée à la dimension géométrique caractéristique du domaine.

Discrétisons (5.6)-(5.8) par rapport à la variable d'espace z. On établit deux systèmes de nœuds : le système de base $\{z_k\}_{k=0}^{N}$, $z_0 = 0$, $z_N = H$, et le système auxiliaire $\{z_{k+1/2}\}_{k=0}^{N-1}$. Les points des deux systèmes alternent, i.e. $z_{k-1/2} < z_k < z_{k+1/2}$.

On intègre la première équation (5.6) par rapport à z dans les limites $(z_0, z_{1/2})$, $(z_{k-1/2}, z_{k+1/2})$, $k = 1, \ldots, N-1$, $(z_{N-1/2}, z_N)$ et la deuxième de z_{k-1} à z_k, $k = 1, \ldots, N$:

$$\frac{\partial}{\partial t}\int_{z_0}^{z_{1/2}} u\,dz + \mu\int_{z_0}^{z_{1/2}} \frac{\partial v}{\partial z}\,dz + \int_{z_0}^{z_{1/2}} \sigma u\,dz =$$

$$= \int_{z_0}^{z_{1/2}} dz \int_0^1 \sigma_s u\,d\mu' + \int_{z_0}^{z_{1/2}} g\,dz,$$

$$\frac{\partial}{\partial t}\int_{z_0}^{z_1} v\,dz + \mu\int_{z_0}^{z_1} \frac{\partial u}{\partial z}\,dz + \int_{z_0}^{z_1} \sigma v\,dz = \int_{z_0}^{z_1} r\,dz,$$

. . . . . .

$$\frac{\partial}{\partial t}\int_{z_{k-1/2}}^{z_{k+1/2}} u\,dz + \mu\int_{z_{k-1/2}}^{z_{k+1/2}} \frac{\partial v}{\partial z}\,dz + \int_{z_{k-1/2}}^{z_{k+1/2}} \sigma u\,dz =$$

$$= \int_{z_{k-1/2}}^{z_{k+1/2}} dz \int_0^1 \sigma_s u\,d\mu' + \int_{z_{k-1/2}}^{z_{k+1/2}} g\,dz, \quad (5.15)$$

$$\frac{\partial}{\partial t}\int_{z_k}^{z_{k+1}} v\,dz + \mu\int_{z_k}^{z_{k+1}} \frac{\partial u}{\partial z}\,dz + \int_{z_k}^{z_{k+1}} \sigma v\,dz = \int_{z_k}^{z_{k+1}} r\,dz,$$

. . . . . .

$$\frac{\partial}{\partial t}\int_{z_{N-1}}^{z_N} v\,dz + \mu\int_{z_{N-1}}^{z_N} \frac{\partial u}{\partial z}\,dz + \int_{z_{N-1}}^{z_N} \sigma v\,dz = \int_{z_{N-1}}^{z_N} r\,dz,$$

$$\frac{\partial}{\partial t}\int_{z_{N-1/2}}^{z_N} u\,dz + \mu\int_{z_{N-1/2}}^{z_N} \frac{\partial v}{\partial z}\,dz + \int_{z_{N-1/2}}^{z_N} \sigma u\,dz =$$

$$= \int_{z_{N-1/2}}^{z_N} dz \int_0^1 \sigma_s u\,d\mu' + \int_{z_{N-1/2}}^{z_N} g\,dz.$$

On introduit les notations

$$\Delta z_0 = z_{1/2} - z_0, \quad \Delta z_k = z_{k+1/2} - z_{k-1/2}, \quad k = 1, \ldots, N-1,$$
$$\Delta z_N = z_N - z_{N-1/2}, \quad \Delta z_{k-1/2} = z_k - z_{k-1}, \quad k = 1, \ldots, N,$$
$$h = \max(\Delta z_0, \Delta z_N, \Delta z_k, \Delta z_{k+1/2}), \tag{5.16}$$

$$\sigma_k = \frac{1}{\Delta z_k} \int_{z_{k-1/2}}^{z_{k+1/2}} \sigma \, dz,$$

$$\sigma_{sk} = \frac{1}{\Delta z_k} \int_{z_{k-1/2}}^{z_{k+1/2}} \sigma_s \, dz, \quad \sigma_{k+1/2} = \frac{1}{\Delta z_{k+1/2}} \int_{z_k}^{z_{k+1}} \sigma \, dz, \tag{5.17}$$

$$g_k = \frac{1}{\Delta z_k} \int_{z_{k-1/2}}^{z_{k+1/2}} g \, dz, \quad r_{k+1/2} = \frac{1}{\Delta z_{k+1/2}} \int_{z_k}^{z_{k+1}} r \, dz.$$

Si u et v sont continues pour presque tous les z et μ de D, les fonctions σ, σ_s, g et r sont continues par morceaux avec des discontinuités éventuelles de première espèce aux points z_k, les méthodes du numéro 2.3 conduisent, compte tenu des conditions aux limites, à l'approximation aux différences des équations (5.6)

$$\begin{aligned}
&\frac{\partial u_0}{\partial t} + \mu \frac{v_{1/2} - u_0}{\Delta z_0} + \sigma_0 u_0 = \sigma_{s,0} \int_0^1 u_0 \, d\mu' + g_0, \\
&\frac{\partial v_{1/2}}{\partial t} + \mu \frac{u_1 - u_0}{\Delta z_{1/2}} + \sigma_{1/2} v_{1/2} = r_{1/2}, \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\frac{\partial u_k}{\partial t} + \mu \frac{v_{k+1/2} - v_{k-1/2}}{\Delta z_k} + \sigma_k u_k = \sigma_{s,k} \int_0^1 u_k \, d\mu' + g_k, \\
&\frac{\partial v_{k+1/2}}{\partial t} + \mu \frac{u_{k+1} - u_k}{\Delta z_{k+1/2}} + \sigma_{k+1/2} v_{k+1/2} = r_{k+1/2}, \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\frac{\partial v_{N-1/2}}{\partial t} + \mu \frac{u_N - u_{N-1}}{\Delta z_{N-1/2}} + \sigma_{N-1/2} v_{N-1/2} = r_{N-1/2}, \\
&\frac{\partial u_N}{\partial t} + \mu \frac{u_N - v_{N-1/2}}{\Delta z_N} + \sigma_N u_N = \sigma_{s,N} \int_0^1 u_N \, d\mu' + g_N.
\end{aligned} \tag{5.18}$$

Appelons $M_h(0, 2N)$ l'espace de Hilbert des fonctions vectorielles $a = (a_0, a_{1/2}, \ldots, a_{N-1/2}, a_N)$ muni du produit scalaire et de

la norme associée

$$(a, b) = \sum_{i=0}^{2N} \int_0^1 \Delta z_{i/2} a_{i/2} b_{i/2}\, d\mu,$$

$$\| a \| = (a, a)^{1/2}, \qquad a, b \in M_h(0, 2N).$$

On introduit les fonctions vectorielles

$$\varphi = (u_0, v_{1/2}, u_1, \ldots, u_{N-1}, v_{N-1/2}, u_N),$$

$$F = (g_0, r_{1/2}, g_1, \ldots, g_{N-1}, r_{N-1/2}, g_N),$$

$$\varphi^{(0)} = (u_0^{(0)}, v_{1/2}^{(0)}, u_1^{(0)}, \ldots, u_{N-1}^{(0)}, v_{N-1}^{(0)}, u_N^{(0)})$$

et l'opérateur matriciel $A = L - S$, où

$$L = \begin{Vmatrix} \frac{\mu}{\Delta z_0} + \sigma_0 & \frac{\mu}{\Delta z_0} & 0 & \ldots & 0 & 0 \\ -\frac{\mu}{\Delta z_{1/2}} & \sigma_{1/2} & \frac{\mu}{\Delta z_{1/2}} & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \sigma_{N-1/2} & \frac{\mu}{\Delta z_{N-1/2}} \\ 0 & 0 & 0 & \ldots & -\frac{\mu}{\Delta z_N} & \frac{\mu}{\Delta z_N} + \sigma_N \end{Vmatrix},$$

$$S = \operatorname{diag}\left(\sigma_{s,\, i/2} \int_0^1 \gamma_{i/2}\, d\mu\right), \qquad i = 0, \ldots, 2N.$$

Ici

$$\gamma_{i/2} = \begin{cases} 1 \text{ si } i/2 \text{ est entier ou } 0, \\ 0 \text{ dans le cas contraire.} \end{cases}$$

Avec ces notations, le système (5.18) s'écrit sous forme opératorielle

$$\frac{d\varphi}{dt} + A\varphi = F, \quad t \in [0, T], \qquad (5.19)$$

$$\varphi = \varphi^{(0)} \quad \text{pour} \quad t = 0.$$

Etablissons certaines propriétés des opérateurs du problème. L'opérateur S est auto-adjoint et positif dans $M_h(0, 2N)$ (propriétés évidentes sous l'hypothèse faite sur les données initiales). On montre que A et L sont définis positifs dans cet espace. Soit $w \in M_h$, auquel

cas on a pour la forme quadratique $(Lw,\ w)$

$$(Lw,\ w)=\int_0^1 d\mu\left(\mu\frac{w_{1/2}+w_0}{\Delta z_0}+\sigma_0 w_0\right)\Delta z_0 w_0+$$

$$+\int_0^1 d\mu\left(\mu\frac{w_1-w_0}{\Delta z_{1/2}}+\sigma_{1/2}w_{1/2}\right)\Delta z_{1/2}w_{1/2}+$$

$$+\sum_{i=1}^{N-1}\int_0^1 d\mu\left\{\Delta z_i\left(\mu\frac{w_{i+1/2}-w_{i-1/2}}{\Delta z_i}+\sigma_i w_i\right)w_i+\right.$$

$$\left.+\Delta z_{i+1/2}\left(\mu\frac{w_{i+1}-w_i}{\Delta z_{i+1/2}}+\sigma_{i+1/2}w_{i+1/2}\right)w_{i+1/2}\right\}+$$

$$+\int_0^1 d\mu\Delta z_N\left(\mu\frac{w_N-w_{N-1/2}}{\Delta z_N}+\sigma_N w_N\right)w_N=$$

$$=\sum_{i=0}^{2N}\int_0^1 d\mu\Delta z_{i/2}\sigma_{i/2}w_{i/2}^2+\int_0^1 d\mu\mu\,(w_0^2+w_N^2),\qquad(5.20)$$

$$(Lw,\ w)\geqslant\gamma\,\|w\|^2,\ \gamma=\text{const}>0,$$

et

$$(Aw,\ w)=\int_0^1 d\mu\mu\,(w_0^2+w_N^2)+\sum_{i=0}^{2N}\Delta z_{i/2}\int_0^1 d\mu\left(\sigma_{i/2}w_{i/2}^2-\right.$$

$$\left.-\sigma_{si/2}w_{i/2}\int_0^1 d\mu\gamma_{i/2}w_{i/2}\right)\geqslant\sigma_{c0}\,\|w\|^2,\qquad(5.21)$$

c.q.f.d.

Etablissons deux estimations à priori de la solution de (5.19). On multiplie (5.19) scalairement par φ et on intègre sur $(0,\ t)$:

$$\frac{1}{2}\,\|\varphi\|^2(t)+\int_0^t dt\,(A\varphi,\ \varphi)=\int_0^t dt^1\,(F,\ \varphi)+\frac{1}{2}\,\|\varphi^{(0)}\|^2\leqslant$$

$$\leqslant\left(\int_0^t dt^1\,\|F\|^2\right)^{1/2}\left(\int_0^t dt^1\,\|\varphi\|^2\right)^{1/2}+\frac{1}{2}\,\|\varphi^{(0)}\|^2\leqslant$$

$$\leqslant C\left(\int_0^t dt^1\,\|F\|^2\right)^{1/2}\left(\frac{1}{2}\,\|\varphi\|^2(t)+\int_0^t dt^1\,(A\varphi,\ \varphi)\right)^{1/2}+\frac{1}{2}\,\|\varphi^{(0)}\|^2,$$

$$C=\text{const}>0.$$

Etant donnée la relation $|a \cdot b| \leqslant \frac{1}{4\varepsilon} a^2 + \varepsilon b^2$, avec $\varepsilon > 0$ choisi convenablement, la dernière inégalité fournit l'une des estimations cherchées :

$$\| \varphi \|^2 (t) + \int_0^t dt^1 (A\varphi, \varphi) \leqslant C \left(\int_0^T dt^1 \| F \|^2 + \| \varphi^{(0)} \|^2 \right), \tag{5.22}$$

la constante $C > 0$ étant indépendante de t et φ.

On démontre moyennant (5.22) l'unicité de la solution de (5.19), l'ordre de régularité en t de φ étant d'une unité supérieur à celui de la fonction vectorielle F. Sans nous arrêter sur cette affirmation, nous supposerons à priori que (5.19) n'a pas d'autre solution que φ possédant la propriété de régularité mentionnée.

Décomposons l'opérateur A de façon que $A = B + D$, où

$$D = \begin{Vmatrix} \frac{\mu}{\Delta z_0} + \sigma_0 & 0 & 0 \ldots & 0 & 0 \\ 0 & \sigma_{1/2} & 0 \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 \ldots & \sigma_{N-1/2} & 0 \\ 0 & 0 & 0 \ldots & 0 & \frac{\mu}{\Delta z_N} + \sigma_N \end{Vmatrix}.$$

L'opérateur D est évidemment défini positif dans $M_h(0, 2N)$. L'équation (5.19) équivaut donc à

$$\frac{d}{dt} (D^{-1} v) + D^{-1/2} A D^{-1/2} v = D^{-1/2} F, \tag{5.23}$$

où $v = D^{1/2}\varphi$. On multiplie (5.23) scalairement par v et on intègre en t de 0 à t_1 :

$$\frac{1}{2} \| \varphi \|^2 (t_1) + \int_0^{t_1} dt \, (D^{-1/2} A D^{-1/2} v, v) = \int_0^{t_1} dt \, (D^{-1/2} F, v) + \frac{1}{2} \| \varphi^{(0)} \|^2,$$

d'où l'on tire par de simples évaluations

$$\frac{1}{2} \| \varphi \|^2 (t_1) + \int_0^{t_1} dt \, (D^{-1/2} \tilde{D} D^{-1/2} v, v) \leqslant$$

$$\leqslant \frac{1}{2} \| \varphi \|^2 (t_1) + \int_0^{t_1} dt \, (D^{-1/2} A D^{-1/2} v, v) \leqslant$$

$$\leqslant \left(\int_0^{t_1} dt \, \| D^{-1/2} F \|^2 \right)^{1/2} \left(\int_0^{t_1} dt \, \| v \|^2 \right)^{1/2} + \frac{1}{2} \| \varphi^{(0)} \|^2, \tag{5.24}$$

avec

$$\widetilde{D}=\left\|\begin{array}{cccc} \frac{\mu}{\Delta z_0}+\sigma_{c,0} & 0 & 0 & 0 \\ 0 & \sigma_{c,1/2} \ldots & 0 & 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & \\ 0 & 0 \ldots & \sigma_{c,N-1/2} & 0 \\ 0 & 0 & 0 & \frac{\mu}{\Delta z_N}+\sigma_{c,N} \end{array}\right\|.$$

Comme

$$\int_0^{t_1} dt\,(D^{-1/2}\widetilde{D}D^{-1/2}v,\ v)\geqslant \int_0^{t_1} dt \int_0^1 d\mu \left(\Delta z_0 \frac{\mu+\Delta z_0\sigma_{c0}}{\mu+\Delta z_0\sigma_0} v_0^2 + \right.$$

$$\left. + \Delta z_{1/2} \frac{\sigma_{c,1/2}}{\sigma_{1/2}} v_{1/2}^2 + \ldots + \Delta z_N \frac{\mu+\Delta z_N\sigma_{c,N}}{\mu+\Delta z_N\sigma_N} v_N^2\right) \geqslant C_1 \int_0^{t_1} dt \,\|v\|^2,$$

avec $C_1 > 0$ une constante indépendante de v et $\Delta z_{i/2}$, la dernière inégalité aidant, on déduit de (5.24) la seconde estimation à priori de φ:

$$\frac{1}{2}\|\varphi\|^2(t_1)+C_1\int_0^{t_1} dt\,\|v\|^2 \leqslant$$

$$\leqslant \left(\int_0^{t_1} dt\,\|D^{-1/2}F\|^2\right)^{1/2}\left(\int_0^{t_1} dt\,\|v\|^2\right)^{1/2}+\frac{1}{2}\|\varphi^{(0)}\|^2, \quad (5.25)$$

$$\frac{1}{2}\|\varphi(t_1)\|^2+\int_0^{t_1} dt\,\|D^{1/2}\varphi\|^2 \leqslant C\left(\int_0^{t_1} dt\,\|D^{-1/2}F\|^2+\|\varphi^{(0)}\|^2\right),$$

où la constante $C > 0$ ne dépend pas de φ et $\Delta z_{i/2}$.

Utilisons cette estimation pour évaluer l'écart entre la solution du problème (5.19) et la fonction vectorielle $\varphi_T = (u(z_0, \mu, t), v(z_{1/2}, \mu, t), \ldots, v(z_{N-1/2}, \mu, t), u(z_N, \mu, t))$ déterminée par les valeurs de la solution exacte de (5.6). On suppose que la solution et les données initiales du problème (5.6)-(5.8) sont suffisamment régulières par rapport à toutes les variables, le pas du réseau est uniforme, i.e. $h = H/N$, $z_i = ih$, $z_{i-1/2} = (i - 1/2)h$, et h est suffisamment faible. Les estimations d'erreur pour cette classe fonctionnelle assez restreinte plaident en la faveur de la méthode d'approximation considérée pour plusieurs problèmes pratiques.

On note que les erreurs d'approximation ε_i, $i = 0, N$, dans la première et la dernière équation (5.18) sont d'ordre 1 en h, les erreurs $\varepsilon_{i/2}$ étant en $O(h^2)$. On remplace φ de (5.25) par $\varphi_T - \varphi$

et F par $\varepsilon = (\varepsilon_0, \varepsilon_{1/2}, \ldots, \varepsilon_{N-1/2}, \varepsilon_N)$ et on obtient les estimations cherchées :

$$\| \varphi_T - \varphi \|^2 (t_1) + \int_0^{t_1} dt \, \| D^{1/2} (\varphi_T - \varphi) \|^2 \leqslant C \int_0^{t_1} dt \, \| D^{-1/2} \varepsilon \|^2 \leqslant$$

$$\leqslant C \int_0^T dt \int_0^1 d\mu \left[\frac{\varepsilon_0^2 h^2}{2(2\mu + h\sigma_0)} + \frac{\varepsilon_{1/2}^2 h}{\sigma_{1/2}} + \ldots + \frac{\varepsilon_{N-1/2}^2 h}{\sigma_{N-1/2}} + \frac{\varepsilon_N^2 h^2}{2(2\mu + h\sigma_N)} \right] \leqslant$$

$$\leqslant O(h^4) + O(h^4) \int_0^1 \frac{d\mu}{2\mu + h\sigma_0} + O(h^4) \int_0^1 \frac{d\mu}{2\mu + h\sigma_N} = O\left(h^4 \ln \frac{1}{h}\right), \quad (5.26)$$

$$\max_t \| \varphi_T - \varphi \| (t) + \left(\int_0^T dt \, \| D^{1/2} (\varphi_T - \varphi) \|^2 \right)^{1/2} = O\left(h^2 \ln^{1/2} \frac{1}{h}\right).$$

Les développements ci-dessus et les estimations à priori sont dus à Agochkov.

Les inégalités (5.25) permettent d'évaluer les erreurs sur la solution de (5.19) avec des contraintes plus lâches sur la régularité des données initiales et de la solution.

Décrivons la méthode de décomposition pour le problème (5.19). On introduit les opérateurs A_1, A_2 agissant dans $M_h(0, 2N)$

$$A_1 = \begin{Vmatrix} \frac{\mu}{\Delta z_0} & \frac{\mu}{\Delta z_0} & 0 & \ldots & 0 & 0 \\ -\frac{\mu}{\Delta z_{1/2}} & 0 & \frac{\mu}{\Delta z_{1/2}} & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & \frac{\mu}{\Delta z_{N-1/2}} \\ 0 & 0 & 0 & \ldots & -\frac{\mu}{\Delta z_N} & \frac{\mu}{\Delta z_N} \end{Vmatrix},$$

$$A_2 = \operatorname{diag}\left(\sigma_{i/2} - \sigma_{s,\, i/2} \int_0^1 d\mu' \gamma_{i/2} \right)$$

tels que

$$A = A_1 + A_2.$$

On montre sans peine la positivité de A_1 et on établit tout aussi facilement à l'aide de

$$\int_0^1 d\mu \left(\sigma_{i/2} a_{i/2}^2 - \sigma_{s,\, i/2} a_{i/2} \int_0^1 d\mu' \gamma_{i/2} a_{i/2} \right) \geqslant \sigma_{c,\, 0} \int_0^1 d\mu \, a_{i/2}^2$$

que A_2 est défini positif dans $M_h(0, 2N)$, i.e.

$$(A_1 a, a) \geqslant 0, \quad (A_2 a, a) \geqslant \gamma \| a \|^2, \quad \gamma = \text{const} > 0.$$

Les conditions de définitivité permettent de formuler un algorithme de splitting-up à deux cycles (voir n° 4.3). Formulons les problèmes suivants :

sur l'intervalle $t_{j-1} \leqslant t \leqslant t_j$

$$\begin{aligned} \left(E + \tfrac{\tau}{2} A_1^h\right) \varphi^{j-2/3} &= \left(E - \tfrac{\tau}{2} A_1^h\right) \varphi^{j-1}, \\ \left(E + \tfrac{\tau}{2} A_2^h\right) \varphi^{j-1/3} &= \left(E - \tfrac{\tau}{2} A_2^h\right) \varphi^{j-2/3}, \end{aligned} \tag{5.27}$$

sur l'intervalle $t_{j-1} \leqslant t \leqslant t_{j+1}$

$$\varphi^{j+1/3} = \varphi^{j-1/3} + 2\tau F^j, \tag{5.28}$$

sur l'intervalle $t_j \leqslant t \leqslant t_{j+1}$

$$\begin{aligned} \left(E + \tfrac{\tau}{2} A_2^h\right) \varphi^{j+2/3} &= \left(E - \tfrac{\tau}{2} A_2^h\right) \varphi^{j+1/3}, \\ \left(E + \tfrac{\tau}{2} A_1^h\right) \varphi^{j+1} &= \left(E - \tfrac{\tau}{2} A_1^h\right) \varphi^{j+2/3}, \end{aligned} \tag{5.29}$$

F^j étant un vecteur de composantes

$$F_i^j = \frac{1}{t_{j+1} - t_{j-1}} \int_{t_{j-1}}^{t_{j+1}} dt F_i, \qquad t_j = j\tau, \quad \tau = T/y.$$

Les résultats du Chapitre premier et les propriétés de A_1, A_2 entraînent que le système (5.27)-(5.29) approche, pour une solution suffisamment régulière, le problème (5.19) à une quantité en $O(\tau^2)$ près et est stable :

$$\max_j \| \varphi^{(j)} \| \leqslant C \left(\| \varphi^{(0)} \| + \max_j \| F^{(j)} \| \right), \tag{5.30}$$

$C = \text{const} > 0$. Ce résultat d'approximation et l'inégalité (5.30) impliquent

$$\max_j \| \varphi(t_j) - \varphi^{(j)} \| \leqslant O(\tau^2). \tag{5.31}$$

Si (5.26) a lieu, la solution de (5.27)-(5.29) converge pour la métrique de $M_h(0, 2N)$ vers la solution exacte du problème initial quand $h, \tau \to 0$, et

$$\max_j \| \varphi_T(t_j) - \varphi^{(j)} \| \leqslant O\left\{ h^2 \left| \ln^{1/2} \frac{1}{h} \right| + \tau^2 \right\}, \tag{5.32}$$

où

$$\varphi_T(t_j) = (u(z_0, t_j, \mu), v(z_{1/2}, \mu, t_j), \ldots, v(z_{N-1/2}, \mu, t_j), u(z_N, \mu, t_j)).$$

Proposons-nous de résoudre le système (5.27)-(5.29). La matrice $\left(E + \frac{\tau}{2} A_1\right)$ étant tridiagonale, on résout facilement pour tout μ fixe la première équation (5.27) et la seconde équation (5.29). L'opérateur matriciel $\left(E + \frac{\tau}{2} A_2\right)$ est diagonal, si bien que la solution de la seconde équation (5.27) est calculée par les formules

$$\varphi_i^{j-1/3} = \frac{1}{1+\frac{\tau\sigma_i}{2}} \left[\left(1 - \frac{\tau\sigma_i}{2}\right) \varphi_i^{j-2/3} + \frac{\tau\sigma_{s,\,i}}{1+\frac{\tau\sigma_{c,\,i}}{2}} \int_0^1 d\mu \varphi_i^{j-2/3} \right],$$

$$i = 0, 1, \ldots, N, \tag{5.33}$$

$$\varphi_{i-1/2}^{j-1/3} = \frac{1-\frac{\tau\sigma_i}{2}}{1+\frac{\tau\sigma_i}{2}} \varphi_{i-1/2}^{j-2/3}, \qquad i = 1, \ldots, N.$$

On procède de même pour la première équation (5.29). Ainsi, on a déterminé l'algorithme numérique pour le système en question.

Discrétisons les équations par rapport à μ. On partage l'intervalle $0 \leqslant \mu \leqslant 1$ par les nœuds μ_l en les intervalles partiels $\Delta\mu_l$ de façon que la classe donnée de solutions fournisse une meilleure approximation des intégrales de (5.33). Soit

$$\int_0^1 d\mu \psi(\mu) \cong \sum_{l=1}^{m} s_l \psi_l, \quad \psi_l = \psi(\mu_l),$$

s_l étant les poids de la formule de quadrature choisie. Remplaçons les intégrales de (5.27)-(5.29) par les formules de quadrature et considérons le système pour $\mu = \mu_l$, $l = 1, \ldots, m$, il vient le système d'équations algébriques linéaires

$$\left(E + \frac{\tau}{2} A_{1,\,l}\right) \varphi_l^{j-2/3} = \left(E - \frac{\tau}{2} A_{1,\,l}\right) \varphi_l^{j-1},$$

$$\left(E + \frac{\tau}{2} A_{2,\,l}\right) \varphi_l^{j-1/3} = \left(E - \frac{\tau}{2} A_{2,\,l}\right) \varphi_l^{j-2/3},$$

$$\varphi_l^{j+1/3} = \varphi_l^{j-1/3} + 2\tau F_l^j, \tag{5.34}$$

$$\left(E + \frac{\tau}{2} A_{2,\,l}\right) \varphi_l^{j+2/3} = \left(E - \frac{\tau}{2} A_{2,\,l}\right) \varphi_l^{j+1/3},$$

$$\left(E + \frac{\tau}{2} A_{1,\,l}\right) \varphi_l^{j+1} = \left(E - \frac{\tau}{2} A_{1,\,l}\right) \varphi_l^{j+2/3},$$

$$\varphi_l^0 = \varphi_l^{(0)},$$

qui approche le problème initial. Ici $\varphi_l^{j-1}, \ldots, \varphi_l^{j+1}$ sont des vecteurs de dimension $2N + 1$, $l = 1, \ldots, m$, $\varphi_l^{(0)} = \varphi^{(0)}(\mu_l)$,

$A_{1,\,l} = A_1(\mu_l)$ et $A_{2,\,l}$ opère selon la formule

$$(A_{2,\,l}\varphi_l^j)_{i/2} = \sigma_{i/2}\varphi_{l,\,i/2}^j - \sigma_{s,\,i/2}\sum_{k=1}^{m} s_k \varphi_{k,\,i/2}^j, \qquad i = 0, \ldots, 2N.$$

On sait déjà qu'on résout facilement la première et la dernière équation (5.34) par exemple par le balayage. Dans les deux cas restants, on utilise les formules

$$\varphi_{l,\,i}^{j-1/3} = \frac{1}{1+\frac{\tau\sigma_i}{2}}\left[\left(1-\frac{\tau\sigma_i}{2}\right)\varphi_{l,\,i}^{j-2/3} + \frac{\tau\sigma_{si}}{1+\frac{\tau\sigma_{ci}}{2}}\sum_{k=1}^{m} s_k \varphi_{k,\,i}^{j-2/3}\right],$$

$$i = 0, \ldots, N, \tag{5.35}$$

$$\varphi_{l,\,i-1/2}^{j-1/3} = \frac{1-\frac{\tau\sigma_i}{2}}{1+\frac{\tau\sigma_i}{2}}\varphi_{l,\,i-1/2}^{j-2/3}, \qquad i = 1, \ldots, N,$$

où l'on fait la substitution $j = j + 1$ s'il s'agit de la quatrième équation.

Ainsi, on a complètement défini l'algorithme numérique pour l'équation de transport non stationnaire qui conduit pour des solutions régulières à un schéma absolument stable en $O(h^2(\ln 1/h)^{1/2} + \tau^2)$. L'ordre de précision en μ dépend de la formule de quadrature choisie. (On note que le schéma considéré admet $\mu_l = 0$ comme point nodal.)

Les paramètres du schéma n'ont pas été soumis à des contraintes. Il y a cependant lieu de dire qu'avec (5.34) abordé par la méthode de factorisation, il suffit que $\tau < \min_i(\Delta z_{i/2})$ pour qu'il y ait stabilité. On conçoit que dans la pratique cette condition restreint le choix du pas de temps.

L'analyse faite se transpose aux équations de transport multidimensionnelles.

CHAPITRE 9

UN APERÇU DES MÉTHODES NUMÉRIQUES

Avec l'avènement des calculateurs électroniques puissants, de nombreux domaines de la science et de la technique sont devenus le siège d'un travail d'algorithmisation intense et d'expériences mathématiques d'envergure, ce qui a accru l'attrait du calcul automatique. La pratique des problèmes d'applications devra être mise à profit pour élaborer les algorithmes et les méthodes numériques efficaces.

Ce chapitre passe brièvement en revue les directions principales du Calcul numérique moderne et leurs lignes d'évolution.

9.1. Approximation, stabilité et convergence des schémas aux différences

La vogue des méthodes des différences finies pour les équations différentielles relevant de la physique mathématique a fait étudier à fond les propriétés des équations aux différences, qui influent directement sur la qualité des schémas discrétisés. Il s'agit avant tout des propriétés de stabilité et de convergence.

La théorie de la stabilité et de la convergence tire ses origines de l'époque où on s'est aperçu, grâce au calcul sur machines rapides, que le schéma aux différences approximant un problème différentiel bien posé peut être instable (mal posé). Or, un schéma instable est sensible aux erreurs d'arrondi, ce qui risque de conduire à un résultat qui diffère catastrophiquement de la solution du problème différentiel.

Ce caractère spécifique des équations aux différences a nourri les recherches visant à établir la relation entre la convergence et la stabilité.

Autour des années 1950, Lax [VI], Lax, Richtmyer [VII], Richtmyer [III] [VI ; 2], Rjabenki, Filippow [VI], von Neumann, Richtmyer [VII] ont énoncé, chacun à sa manière, un résultat fondamental connu sous le nom de théorème d'équivalence : étant donné un problème différentiel linéaire homogène, s'il est bien posé et appro-

ché par un schéma aux différences, alors la stabilité du schéma est une condition suffisante pour que la solution du problème aux différences converge vers celle du problème différentiel. Rjabenki, Filippow [VI] l'ont formulé et démontré pour une équation abstraite dans un espace de Banach, et Richtmyer [III] l'a généralisé au cas d'équations différentielles linéaires non homogènes. Le théorème d'équivalence est énoncé en termes de même norme. La convergence en d'autres normes est de règle établie par les théorèmes de l'immersion de Sobolev [I ; 6]. Si l'on renforce les conditions de régularité imposées aux données initiales, on peut affaiblir les contraintes sur la stabilité du schéma (cette circonstance a été notée pour la première fois par Rjabenki, Filippow [VI]). Strang [VII] asseoit dessus son théorème d'équivalence qui utilise la notion de stabilité faible.

En ce qui concerne les critères efficaces de stabilité, signalons en premier lieu le travail de von Neumann et Richtmyer [VII]. Comme le critère local de ces auteurs ne joue que pour les équations à coefficients constants et les problèmes auto-adjoints, on a voulu en localiser le domaine d'action.

Lax et Nirenberg [VI] ont élaboré une théorie de la stabilité des schémas aux différences hyperboliques en termes de symbole de l'opérateur discrétisé. Dans le cas d'équations explicites, le symbole coïncide avec la matrice de transition classique obtenue par la méthode de Fourier, et le critère local de stabilité est valable si les coefficients possèdent des dérivées secondes bornées par rapport à x.

Strang [VII] a énoncé un théorème de convergence pour les systèmes d'équations hyperboliques quasi linéaires sous l'hypothèse de stabilité locale des équations aux différences associées à la variation première du système différentiel et de régularité suffisante de la solution.

L'étude des schémas aux différences à coefficients variables utilise la notion de dissipativité. Signalons les théorèmes dus à Kreiss [VI] qui établissent la relation entre l'ordre de dissipativité et l'ordre de précision des équations aux différences qui approchent les systèmes hyperboliques, les coefficients matriciels des approximations étant supposés être des fonctions hermitiennes et continues lipschitziennes de x.

La stabilité est traitée de façon intéressante dans Yanenko, Chokin [VII ; 8] : au lieu de l'équation aux différences, on considère une équation différentielle associée appelée première approximation différentielle (PAD) dont la propriété d'être mal posée entraîne l'instabilité du schéma aux différences.

Une classe fort importante est constituée des schémas aux différences à opérateurs positifs étudiés par Friedrichs [II]. Cet auteur a introduit la notion générale de schéma positif et établi une condition suffisante de stabilité dans L_2.

On doit à Godounov, Riabenki [VII ; 2] la notion de spectre d'une famille d'opérateurs aux différences et une condition nécessaire de stabilité qui s'inspire de cette notion et traduit bien le phénomène d'instabilité. Ces auteurs ont également introduit la notion de noyau dudit spectre. Les normes des puissances des opérateurs sont évaluées en termes de rayons des noyaux, et les estimations s'avèrent uniformes pour toute la famille et susceptibles de trancher la question de la stabilité.

Les critères de stabilité considérés sont dits spectraux parce que procédant du spectre d'opérateurs discrets. Ils permettent d'établir la convergence en norme de L_2. La stabilité dans C est démontrée par Serdukova [VI ; 4], Thomée [VI], Samarski [VII ; 5].

Parmi les approches non spectrales de la stabilité des équations hyperboliques et paraboliques discrètes, la théorie très générale de Samarski [III ; 12] [VI ; 3] [VII ; 5] s'appuie sur les inégalités d'énergie et les estimations à priori. Elle donne, pour une vaste classe de schémas à deux et à trois niveaux, des conditions suffisantes de stabilité sous forme d'inégalités entre les coefficients opératoriels. Ces conditions sont fort constructives et permettent non seulement de tester la stabilité, mais aussi de former des schémas stables.

On assiste à un développement continu de la méthode de l'énergie qui consiste à chercher une norme du vecteur solution dont la valeur augmente à chaque pas au plus comme $1 + O(\Delta t)$, ce qui signifie la stabilité pour cette norme.

La méthode remonte à Courant, Friedrichs, Lewy [VII]. Elle a été développée avec succès dans Ladyjenskaïa [VII ; 3], Lees [VII], Lax [VI], Kreiss [VI], Samarski [III ; 12], Konovalov [XV ; 8].

S'agissant des problèmes aux limites hyperboliques, la stabilité a été étudiée par Kreiss [VI] qui a établi par exemple des conditions suffisantes de stabilité des problèmes discrétisés sous des hypothèses assez lâches sur les données initiales.

Kantorovich, Akilov [I] ont développé la théorie de l'approximation et de la convergence à la lumière de l'Analyse fonctionnelle. Ils ont considéré une classe étendue d'équations opératorielles tout en insistant sur la résolution numérique d'équations intégrales.

Un grand rôle incombe en théorie de la convergence au concept de fermeture des algorithmes de calcul de Sobolev [I ; 6]. On utilise volontiers ce concept pour légitimer les méthodes d'approximation en physique mathématique.

9.2. Méthodes numériques pour les problèmes de la physique mathématique

Avec les notions d'approximation, de stabilité et de convergence, on pouvait s'attaquer en toute sécurité à la recherche des schémas aux différences efficaces pour les problèmes de la physique mathé-

matique. Les algorithmes de différences finies unissent en général des procédés de discrétisation à des méthodes de résolution des problèmes ainsi obtenus. Si la théorie constructive des méthodes des différences enregistre des succès, elle le doit donc au développement proportionné des deux directions indiquées.

La théorie et la pratique des méthodes par différences finies se sont enrichies ces dernières années de résultats nombreux et significatifs qui autorisent à faire le point et à dégager les aspects fondamentaux de la question.

CONSTRUCTION DE SCHÉMAS AUX DIFFÉRENCES. On élabore les méthodes de construction de schémas aux différences qui expriment les lois de conservation propres à la plupart des phénomènes physiques. Ces schémas sont formés moyennant les équations du bilan écrites pour chaque maille du réseau et les formules de quadrature et d'interpolation. Après avoir effectué certaines transformations et sommé par rapport à tous les points du réseau, on obtient les équations qui satisfont aux analogues discrets des lois de conservation sous forme intégrale.

Ladyjenskaïa [VII ; 3] a par exemple obtenu, pour les équations à coefficients discontinus, les opérateurs aux différences dont la forme ne change pas d'un point intérieur du réseau à un autre. L'algorithme est justifié à l'aide de la notion de solution généralisée, et on a démontré que la solution du problème aux différences forme une fonctionnelle qui devient, quand $h \to 0$, la fonctionnelle du problème différentiel.

Godounov [IV ; 5], Lax, Wendroff [VI] ont utilisé des approximations aux différences explicites pour construire des schémas conservatifs uniformes de la dynamique des fluides. Il convient de citer la méthode des relations intégrales de Dorodnitsyne [III ; 5] qui a été développée entre autres par Biélotserkovski et Tchouchkine [IV ; 1] ; on discrétise partiellement les équations écrites sous forme de lois de conservation en recourant à la méthode de droites. Ces procédés ont fortement marqué la vision globale de la formation de schémas aux différences pour les équations quasi linéaires. On trouve des concepts généraux intéressants relatifs à l'intégration des équations de la dynamique des fluides dans Babenko et Roussanov (voir Roussanov [XII ; 10], Fromm [IV], Crowley [IV], Kouropatenko [IV ; 13]).

Depuis quelque temps on se penche sur le problème de solutions très précises. Deux tendances sont à noter. La première concerne l'ordre d'approximation élevé des équations discrétisées (voir Samarski [III ; 12], Valioulline, Yanenko [IV ; 2], Mitchell [III]) et la seconde consiste à faire le calcul pour plusieurs pas différents et à prendre des équations aux différences d'ordre d'approximation faible (extrapolation au sens de Richardson). Pour les dernières

méthodes, nous renvoyons à Volkov [IV; 4], Fox, Mayers [III], Chaïdourov [IV; 30].

La technique intégrale d'interpolation a permis à Tikhonov, Samarski [IV; 27] et à d'autres mathématiciens de mettre au point des procédés de construction de schémas aux différences pour les équations elliptiques et paraboliques à coefficients discontinus.

MÉTHODES VARIATIONNELLES. On constate un intérêt accru pour les méthodes procédant du Calcul Variationnel. Les méthodes variationnelles de Ritz, Galerkin et Trefftz occupent depuis des années une place de premier plan en Analyse numérique. Leur efficacité est particulièrement grande dans les problèmes où les inconnues sont des fonctionnelles de la solution. Il se trouve que les approximations d'ordre peu élevé donnent des résultats très précis. Une étude théorique parfaite de ces techniques est due à Mikhlin [I] qui en a établi des conditions nécessaires et suffisantes de stabilité dans les espaces munis de la norme énergie. La pratique des méthodes variationnelles en a révélé des inconvénients : on a à construire des fonctions d'essai qui, même peu nombreuses, fournissent une bonne approximation et tiennent compte des singularités de la solution.

Une nouvelle voie de recherche a été ouverte par les procédés variationnels opérant avec les fonctions d'essai nulles partout, sauf dans une région assez petite du domaine de définition de la solution. Cette direction a débuté avec Courant [V], Oganessian [V; 6], Lions et Temam [V], Céa [V], Aubin [V], Birkhoff, Schultz, Varga [V], Bramble [V]. Ses progrès sont en premier lieu l'œuvre de Babuška [V], Strang, Fix [V], Zlamal [V], Douglas, Dupont [V], Rivkind [IV; 22].

PROBLÈMES STATIONNAIRES À PLUSIEURS VARIABLES. Les méthodes de résolution des équations algébriques linéaires à matrices jacobiennes et tridiagonales par blocs donnent naissance, en matière des problèmes stationnaires, à plusieurs algorithmes numériques de qualité qui reposent sur la factorisation de l'opérateur discrétisé du problème. Parmi les procédés de factorisation, il convient de citer en premier lieu diverses variantes matricielles sans itérations dues à Keldych [V; 3], Guelfand, Lokoutsievski [XII; 5], Babenko, Roussanov (voir Roussanov [XII; 10]), Godounov [XII; 7], Abramov, Andréev [XII; 1]).

Fort des résultats de Vishik, Sobolev et Ljusternik, Sauliev a proposé, pour les problèmes aux limites dans des domaines de géométrie complexe, la méthode du domaine auxiliaire [IV; 25] qui devra être étudiée par Mignot [IV], Koptchénov [IV; 10], Roukhovets [IV; 23], Konovalov [IV; 9]. Lébédev [IV; 15] et Rivkind [IV; 22] se sont par exemple penchés sur les équations aux différences qui approchent un problème aux limites perturbé au sens de la méthode du domaine auxiliaire.

On assiste à des progrès notables des méthodes sans itérations pour les équations aux différences associées à des problèmes diffé-

rentiels à variables séparables. Grâce à la transformation de Fourier rapide et à la réduction cyclique, on réduit sensiblement le volume de calcul (voir Cooley, Tukey [XIII], Buzbee, Golub, Nielson [XII], Hockney [XIII]).

La factorisation approchée fait des progrès aussi rapides que la factorisation proprement dite (cette méthode unit la décomposition de l'opérateur aux approximations successives). La nécessité de ces algorithmes s'est fait sentir en même temps qu'on réussissait à réduire les problèmes de la physique mathématique à des systèmes algébriques de dimension élevée. Les travaux de Bouléev [XV; 3] et de Baker, Oliphant [XV] ont préparé la voie au développement des méthodes qui procèdent des processus rapidement convergents.

Au début des années 1960, l'Analyse numérique s'est enrichie de la méthode des directions alternées de Douglas, Peaceman, Rachford (voir [XV]). La méthode s'est avérée fort séduisante du fait de réduire aisément un problème à plusieurs variables à une succession de problèmes unidimensionnels à matrices du type Jacobi qui s'inversent facilement sur ordinateur. Elle se ramène en fin de compte à une méthode itérative dans laquelle on optimise les calculs par un choix spécial de l'opérateur contractant, produit d'opérateurs plus simples par des paramètres de relaxation libres. L'inversion des opérateurs élémentaires successifs s'effectue en général à l'aide de la factorisation linéaire. Ces schémas itératifs sont très économiques et efficaces, le coût par itération étant légèrement supérieur à celui de la méthode de Richardson explicite. La méthode des directions alternées a beaucoup influencé l'algorithmisation de divers chapitres des mathématiques appliquées et les travaux concernant les processus non locaux et itératifs par blocs. Ses aspects théoriques ont intéressé de nombreux auteurs dont Douglas, Rachford [XV], Diakonov [XV; 5], Samarski [III; 12], Birkhoff, Varga, Young [XV], Wachspress [XV], Kellogg [XV], Gunn [XV], Vorobiev [XV; 4]. Les directions alternées ont éte appliquées par Peaceman, Rachford [XV] à des problèmes non stationnaires.

Les approximations homogènes et non homogènes donnent lieu à des méthodes nouvelles. Dans le cas non homogène, tout problème auxiliaire n'approche pas le problème initial, mais il y a approximation globale pour des normes spéciales. Il s'agit des méthodes de décomposition développées par Yanenko [III; 17, 50], [XV; 13], Diakonov [III; 6] [XV; 5], Samarski [III; 12] [XV; 11], Sauliev [III; 15], Marchuk [XV].

Le problème de choisir les paramètres optimaux des schémas décomposés à l'aide des méthodes spectrales et variationnelles a fait l'objet de nombreux ouvrages (voir p. ex. Samarski [III; 12] [XV; 11], Diakonov [XV; 5], Iline [III; 7] [XV; 7]).

Pour divers aspects théoriques des directions alternées et de la décomposition, voir Andréev [XV; 1], Widlund [XV], Fairweather, Mitchell [XV].

PROBLÈMES NON STATIONNAIRES À PLUSIEURS VARIABLES. La pratique des problèmes unidimensionnels a préparé le terrain pour les algorithmes valables dans des cas plus complexes. Une nouvelle étape dans l'histoire des méthodes de résolution des problèmes non stationnaires en dimension 2 est franchie avec la technique des directions alternées basée sur l'approximation homogène ; primitivement une méthode pour les équations paraboliques à n variables, elle s'est généralisée à nombre d'autres cas.

On sait l'intérêt tout particulier des méthodes de décomposition qui reposent en général sur les approximations aux différences non homogènes. Leur idée de base est de réduire un opérateur complexe à des opérateurs élémentaires, auquel cas on intègre plusieurs équations plus simples que l'équation donnée. Ceci étant, les schémas aux différences vérifient globalement les conditions d'approximation et de stabilité, ce qui permet de former des schémas pour la totalité, ou peu s'en faut, des grands problèmes de la physique mathématique. Divers schémas de décomposition ont été proposés par Bagrinovski, Godounov [XV; 2] (cas explicite) et par Yanenko [XV; 13], Diakonov [XV; 5], Samarski [XV; 11], Marchuk [III; 43 a)] (cas implicite). Ces méthodes s'appliquent avec succès à des problèmes des plus variés. Elles sont à l'origine d'une tactique plus générale basée sur la méthode d'approximation faible due à Yanenko [III; 17, 56] [XV; 13 д)] et à Samarski [III; 12] [XV; 11]. Il se trouve qu'on peut interpréter la décomposition comme l'approximation faible de l'équation initiale par une équation plus simple. Les conditions de convergence sont énoncées par le théorème de Yanenko-Démidov [XV; 14] et dans Lébédev [IV; 15 a), б)] et Diakonov [III; 6] [XV; 5]. La méthode d'approximation faible trouve de nombreuses applications naturelles en dynamique des fluides, en météorologie, en océanographie, en théorie du transport, etc. (voir Marchuk [III; 43 a)] [XVII; 10, 27], Yanenko [III; 17, 56]).

Les problèmes relevant de la dynamique des fluides, de la météorologie et de l'océanographie font volontiers appel au schéma prédicteur-correcteur original de Lax-Wendroff, le prédicteur étant un schéma aux différences explicite. Ce schéma conditionnellement stable est de mise en œuvre facile et approche à l'ordre 2 par rapport à toutes les variables. Pour l'étude du schéma, consulter Richtmyer, Morton [III].

Plusieurs variantes de la prédiction-correction utilisant les approximations aux différences implicites sont proposées par Bryan [IV], Douglas, Rachford [XV], Sofronov [XII; 11], Marchouk, Yanenko [XV; 10]. Il se trouve que tous ces schémas sont

équivalents dans un sens et que seules leurs mises en œuvre diffèrent. Deux derniers auteurs prennent pour prédicteur un schéma décomposé implicite exact à l'ordre 1 à opérateur factorisé. En dynamique des fluides, on utilise en qualité de prédicteurs des schémas majorants implicites.

On doit à Lions et Temam [V], à Bensoussan, Lions et Temam [III] la méthode de décomposition et de décentralisation, un procédé du plus haut intérêt qui s'apparente à la décomposition décrite et à l'approximation faible.

MÉTHODE PIC (THE PARTICLE-IN-CELL METHOD). A l'heure actuelle, on voit augmenter le prestige de la méthode de grosses particules qui relève du procédé connu de Harlow [XIX]. On s'en sert avec succès pour calculer les écoulements fluides à plusieurs dimensions en présence d'une déformation notable de la substance, de déplacements mutuels importants et d'interfaces en collision. Dans cette méthode, on réduit les équations par approximation faible sur chaque intervalle temporel petit à deux systèmes plus simples dont l'un gouverne l'adaptation mutuelle des champs hydrodynamiques, abstraction faite des termes advectifs et est intégré de façon classique dans le réseau fixe d'Euler, et l'autre décrit le transport de substance dans le système de coordonnées de Lagrange. C'est pour résoudre le second système qu'on utilise une simplification phénoménologique du modèle de milieu continu : on le remplace par un ensemble de particules dans chaque cellule du système d'Euler de sorte que le bilan de masse, d'impulsion et d'énergie de cet ensemble s'identifie aux caractéristiques correspondantes du milieu continu. Une particule « porteuse » d'une certaine masse, qui pénètre suivant sa trajectoire propre dans une cellule, apporte avec elle sa masse, son impulsion et son énergie qui sont donc « enlevées » à la cellule qu'elle vient de quitter. Le schéma de Harlow se base dans les deux étapes sur des méthodes explicites. Il est conditionnellement stable globalement. Les schémas implicites employés dans la première phase donnent des résultats particulièrement bons, et le critère de stabilité globale coïncide alors avec la condition connue de Courant. On n'a pas obtenu jusqu'à ce jour de schémas absolument stables, mais il y a lieu d'escompter des progrès dans l'avenir immédiat.

Diatchenko [XIX; 3], Biélotserkovski, Davydov [XIX; 1], Yanenko, Anoutchina et al. [XIX; 4] décrivent des variantes avec fluctuations de densité et de pression beaucoup plus petites et une grande « réserve » de stabilité et examinent divers schémas de mise en œuvre.

On espère qu'avec les méthodes absolument stables et l'élimination des fluctuations, la méthode se généralise aux écoulements peu compressibles et que dans les années à venir ell. entraîne dans son orbite de nombreux problèmes à plusieurs variablese

Voilà plus de deux décennies que la MÉTHODE DE MONTE-CARLO proposée par von Neumann et Ulam connaît une belle fortune. Portée aux nues à ses débuts, elle a été ensuite mise à l'index. Car on n'a pas tardé à s'apercevoir que la condition *sine qua non* de son efficacité est un calculateur effectuant des millions d'opérations par seconde. En effet, on prend dans cette méthode beaucoup d'échantillons qui réduisent l'erreur quadratique moyenne sur le résultat.

Malgré les difficultés qu'on avait à exécuter la méthode sur les machines à vitesse moyenne (et peut-être à cause de ces difficultés), on en a perfectionné la théorie et, partant, l'efficacité pour toute une gamme de problèmes de la science et de la technique. Les perfectionnements les plus sérieux tiennent au fait d'utiliser les probabilités conditionnelles et les poids statistiques définis à l'aide de l'information sur les solutions des équations adjointes par rapport aux fonctionnelles principales des problèmes. Cette façon d'agir réduit de deux et quelquefois de quatre ordres la variance d'erreur et diminue d'autant le temps de calcul par rapport aux méthodes de simulation statistique directe.

Avec la venue des calculateurs de la troisième génération, la méthode de Monte-Carlo s'applique avec succès à divers problèmes complexes de la physique mathématique. Elle jouit d'une bonne réputation dans de nombreux domaines: théorie du transfert de rayonnement, phénomènes d'attente, interpolation, formules de cubature, équations intégrales, systèmes algébriques. Depuis quelque temps, on l'utilise pour les équations de Boltzmann non linéaires et en programmation linéaire.

La théorie et l'algorithmisation de la méthode de Monte-Carlo doivent beaucoup à Vladimirov, Sobol [XVIII ; 2], Guelfand, Frolov, Tchentsov [XVIII ; 3], Fano, Spencer, Berger [XVIII], Ermakov, Zolotoukhine [XVIII ; 5], Mikhaïlov [XVIII ; 9], Bouslenko, Golenko et al. [XVIII ; 1]. Son caractère simple et universel lui assure un brillant avenir en calcul numérique.

9.3. Problèmes conditionnellement bien posés

La résolution numérique d'un problème de la physique mathématique dépend pour une large part de sa propriété d'être « bien ou mal posé », comme disait autrefois l'illustre auteur. Il existe une quantité de problèmes classiques bien posés au sens d'Hadamard. Depuis qu'on a introduit le concept en question, on a eu à faire face à divers problèmes des sciences de la nature et de la technique dont on peut dire qu'ils sont conditionnellement bien posés. Tikhonov [XVI ; 17] a formulé des conditions naturelles aux problèmes mal posés selon Hadamard. Ces conditions complémentaires sont qu'on suppose à priori l'existence d'une solution et son

appartenance à un compact donné. La propriété « le problème est conditionnellement bien posé » est établie en démontrant un théorème d'unicité.

Les problèmes de ce type sont étudiés à fond par Lavrentiev [XVI; 5, 25] et Ivanov [XVI; 1]. Divers aspects de la théorie sont examinés dans John [XVI], Merghélian [XVI; 15], Douglas [XVI], Krein [XVI; 3].

Tikhonov a introduit la notion de régularisation [XVI; 17]: au lieu de l'opérateur non borné qui donne la formule exacte de la solution du problème mal posé, on considère une suite (une famille régularisante) d'opérateurs continus telle qu'il y ait convergence sur chaque élément du domaine d'existence de la solution.

Un moyen original d'aborder les problèmes mal posés au sens d'Hadamard est d'utiliser les notions et les méthodes du Calcul des Probabilités. (La question est particulièrement bien étudiée par Lavrentiev et Vassiliev [XVI; 6].) Les travaux correspondants définissent la notion de stabilité et présentent des algorithmes, optimaux dans un sens, destinés à diverses classes de problèmes sous certaines hypothèses sur les propriétés en probabilité des erreurs de données et de l'ensemble des solutions cherchées.

On doit à Lattès et Lions [XVI] une méthode de quasi-réversibilité (Q.R.) pour les problèmes d'évolution inverses. L'idée de Q.R. consiste à introduire dans l'équation d'évolution l'opérateur régularisant avec paramètre petit qui est le produit de l'opérateur du problème proposé par son adjoint. On choisit ce paramètre par le jeu d'évaluations optimales spéciales. La méthode est de mise en œuvre simple.

L'auteur et Atanbaev [XVI; 11] ont proposé de résoudre des problèmes d'évolution conditionnellement bien posés en utilisant la technique du plus petit résidu et tout le domaine de définition de la solution. On régularise en choisissant le nombre optimal de pas d'itération en fonction de l'estimation à priori des erreurs de données.

On trouve dans Morozov [XIV; 18] [XVI; 16] une étude très complète de la théorie des problèmes mal posés et des méthodes de régularisation.

Il est dans la ligne d'évolution des méthodes de résolution des problèmes conditionnellement bien posés qu'elles s'apparentent aux procédés d'optimisation des processus de calcul.

9.4. Méthodes numériques d'algèbre linéaire

On s'adresse de plus en plus souvent aux grands systèmes d'équations algébriques linéaires à matrices creuses ou pleines, aux systèmes mal conditionnés et aux problèmes spectraux pour les matrices de forme quelconque dont la résolution s'inspire pour une large

part de l'information à priori et à posteriori. Devant la poussée des calculateurs électroniques qui suggèrent des algorithmes nouveaux, commodes en calcul automatique, on s'est vu obligé de revoir les méthodes classiques d'algèbre linéaire.

MÉTHODES DIRECTES. On entend en général par méthode directe d'algèbre linéaire une méthode qui conduit à la solution en un nombre fini d'opérations arithmétiques. On y recourt volontiers dans le cas de systèmes linéaires, de matrices inverses et de déterminants. Ces méthodes mettent une matrice par des transformations élémentaires sous forme de produit de deux matrices simples facilement inversibles.

On range au nombre des méthodes directes classiques l'élimination de Gauss, les méthodes de rotation et de réflexion. Un autre groupe est constitué des techniques utilisant les directions conjuguées (gradient conjugué de Hestenes et Stiefel [XI], itérations minimisées de Lanczos [III]). Les travaux de ces auteurs sont à l'origine des méthodes procédant de l'orthogonalisation.

Ces dernières années, les méthodes directes ont pris de l'ampleur grâce à Faddeev, Faddeeva, Koublanovskaïa [VIII ; 17], Bauer, Fike [VIII], Householder [III], Wilkinson [VIII], Henrici [III], Forsythe, Moler [VIII], Buzbee, Golub, Nielson [XII], Voévodine [VIII ; 3].

Un des problèmes ouverts est la résolution des systèmes à matrices mal conditionnées qui est intimement liée à celle des problèmes conditionnellement bien posés de la physique mathématique. La difficulté tient à la sensibilité de la solution à la précision avec laquelle on se donne les éléments de la matrice et les composantes du vecteur second membre. On n'a obtenu pour le moment que quelques résultats de valeur dans le long chemin qui devra aboutir à la théorie générale.

MÉTHODES ITÉRATIVES. Ces méthodes rendent des services exceptionnels à l'algèbre linéaire. On en connaît plusieurs algorithmes performants facilement exécutables sur ordinateur. Les méthodes itératives se développent essentiellement selon les besoins de la physique mathématique, de l'économie et de la gestion qui se posent des problèmes conduisant à de grands systèmes à matrices de forme spéciale. Le fait est que les méthodes directes s'avèrent peu efficaces dans la plupart de ces problèmes encore que leurs possibilités augmentent avec le perfectionnement du matériel de calcul.

En ce qui concerne la construction des méthodes itératives, plusieurs tendances se font jour actuellement dont nous ne considérerons que deux. La première tendance est d'utiliser les caractéristiques spectrales des opérateurs impliqués, à savoir on génère les itérations où la matrice de transition dépend de l'ensemble des paramètres choisis de deux manières : ou bien on les prend les mêmes à chaque pas en minimisation du rayon spectral de la matrice de

transition, ou bien on en forme une suite dont les termes dépendent du numéro d'itération de façon que le vecteur erreur tende le plus vite possible vers 0 uniformément par rapport à toutes les approximations initiales. Les deux procédés utilisent l'information à priori sur les spectres des matrices. Le choix de ces paramètres constitue une partie intégrante du problème d'optimisation de l'algorithme de calcul, et son point le plus délicat est en général la localisation des spectres des matrices impliquées.

Les progrès sérieux de l'optimisation spectrale posent des problèmes nouveaux. Chose importante, les méthodes spectrales sont particulièrement efficaces pour les problèmes successifs à opérateurs identiques et à entrées différentes.

La deuxième tendance s'inspire des principes variationnels. Les méthodes de cette classe consistent à minimiser de proche en proche une fonctionnelle (en général quadratique) qui atteint enfin son minimum sur la solution cherchée du système. Cette tendance date entre autres de Kantorovitch [XI ; 4], Lanczos [III], Hestenes, Stiefel [XI], Krasnosselski, Krein [XI ; 5]. Parmi les travaux plus récents, il convient de citer Petryshyn [IX] [X], Forsythe [XI], Daniel [XI], Marchouk, Kuznétsov [III] [XI ; 7], Godounov, Prokopov [XI ; 2], Lébédev [IX ; 8], Gorbenko, Iline [XI ; 3].

Les méthodes variationnelles du type plus grande pente et les itérations avec le plus petit résidu ont l'avantage de choisir les paramètres de relaxation à la lumière de l'information à posteriori obtenue à chaque pas. La vitesse de convergence de ces méthodes à pas liés est au moins égale à celle des procédés utilisant les polynômes de Tchebycheff. Leur autre intérêt est de converger pour toute matrice définie positive, symétrique ou non. Le concept de plus petit résidu a dernièrement servi de base à plusieurs méthodes efficaces pour les matrices semi-définies positives.

L'inconvénient sérieux des méthodes variationnelles non stationnaires, qui en entrave d'ailleurs le développement, est la nécessité de mettre en mémoire une plus grande quantité d'information intermédiaire par rapport à l'optimisation au sens de Tchebycheff.

On voit apparaître des méthodes itératives hybrides qui marient les deux tendances. Lébédev a énoncé les conditions à imposer aux opérateurs pour lesquels le processus itératif ne saurait être amélioré en ce qui concerne l'estimation du nombre d'opérations arithmétiques. Un autre moyen de choisir les paramètres optimaux procède du Calcul des Probabilités. Plusieurs résultats intéressants à ce sujet sont dus à Vorobiev [IX ; 3]. La surrelaxation classique de Young-Frankel n'a rien perdu de son importance (voir [X]). Les monographies de Forsythe, Wasow [III], Varga [III], Isaacson, Keller [III], Young [X] font la synthèse des différents travaux consacrés à cette méthode. Pour un aperçu et une classification des méthodes par itérations, voir Marchouk, Lébédev [XVII ; 9].

Les méthodes itératives pour les systèmes linéaires à matrices singulières ont fait l'objet de nombreux ouvrages. S'agissant des systèmes compatibles, l'auteur et Kuznétsov [VIII ; 10, 12, 29] [XI ; 7] ont proposé une tactique générale en vue d'étudier la convergence des méthodes itératives, stationnaires ou non. Outre qu'elle élargit le champ des applications des procédés connus, cette tactique permet d'élaborer la classe de méthodes matricielles du domaine auxiliaire (voir Kouznétsov, Matsokine [IV ; 11]). Quant aux systèmes incompatibles, on trouve des procédés itératifs correspondants dans Moltchanov, Yakovlev [VIII ; 13] et Kuznétsov [VIII ; 9, 28].

Arrêtons-nous sur les méthodes itératives destinées à résoudre complètement le problème de valeurs propres pour les matrices de forme générale. Seules les méthodes des transmutations nous intéressent car c'est dans cette voie qu'on enregistre depuis quelque temps les résultats de valeur dus entre autres à Wilkinson [VIII], Bauer, Fike [VIII], Collatz [III], Voévodine [VIII ; 36)], Frencis [VIII], Koublanovskaïa [VIII ; 8], Eberlein [VIII].

Les méthodes des transmutations consistent à réduire de proche en proche la matrice initiale par des similitudes unitaires (Jacobi, algorithme QR) ou par des similitudes à matrices triangulaires (algorithmes LR) à une matrice dont les valeurs propres sont directement accessibles (ce sont les matrices diagonales, triangulaires ou triangulaires par blocs dont les blocs diagonaux sont d'ordre 2 au plus).

Hier encore, les algorithmes efficaces n'existaient que pour les matrices symétriques (méthode de Jacobi) (Faddeev, Faddeeva [VIII], Rutishauser [VIII], Wilkinson [VIII]). Avec la procédure QR (Koublanovskaïa [VIII ; 8], Frencis [VIII], Wilkinson [VIII]) et la méthode des rotations généralisée (Voévodine [VIII ; 36)]), on aborde le problème de valeurs propres pour les matrices de forme quelconque. Ce sont diverses variantes QR qui tentent le plus les chercheurs.

Le problème considéré est également redevable de ses progrès au calcul des réacteurs nucléaires, car les recherches associées ont imprimé une impulsion aux méthodes itératives pour le problème particulier de valeurs propres en cas des matrices non négatives. Les fondements de la théorie ont été posés par Perron et Frobenius et développés de façon appréciable par Varga [III], Traub [VIII], Marek [VIII ; 11, 30].

Analyse des erreurs d'arrondi. Le temps n'est pas loin où les seuls critères pour comparer deux méthodes numériques étaient le nombre d'opérations arithmétiques et l'occupation mémoire. On y ajoute aujourd'hui la précision, i.e. l'analyse des erreurs d'arrondi commises au cours du calcul sur ordinateur fait partie intégrante des algorithmes. C'est von Neumann qui a ouvert ce chapitre du Calcul numérique, et l'étude systématique des erreurs date de

Wilkinson [VIII]. L'appareil mathématique de Wilkinson repose sur le concept de perturbations équivalentes qui a permis d'évaluer les normes des perturbations pour toutes les transformations d'algèbre linéaire et les normes des perturbations équivalentes pour de nombreuses méthodes.

On assiste à l'essor de la théorie statistique des erreurs. Les résultats de Bakhvalov [VIII ; 2], Voévodine [VIII ; 3], Kim [VIII ; 7] et d'autres auteurs stimulent l'étude de la répartition effective d'erreurs d'arrondi.

ENSEMBLES DE PROGRAMMES STANDARDS. Les progrès réalisés dans les méthodes numériques d'algèbre linéaire ont rendu possible la confection de programmes standards performants pour les systèmes d'équations linéaires et le problème de valeurs propres. La revue *Numerische Mathematik* publie par exemple une quantité de procédures qu'on utilise fréquemment aussi bien dans les problèmes généraux d'algèbre linéaire que dans des problèmes spéciaux à la physique mathématique, à l'économie, etc., qui font intervenir des matrices de forme particulière.

La question ne manquera pas d'intéresser les chercheurs dont les efforts aboutiront à un système de calcul universel pour les problèmes d'algèbre linéaire. On constate au moins deux tendances. L'une est de mettre au point les ensembles de programmes et d'algorithmes pour les cas généraux et l'autre consiste à élaborer des méthodes passe-partout susceptibles de s'adapter à tel problème concret. On comprend l'intérêt capital de ces deux tendances qui préparent la voie au software universel pour les ordinateurs de la quatrième génération et des générations suivantes.

9.5. Optimisation des méthodes numériques

Une des grandes missions de l'Analyse numérique est d'élaborer des procédés de calcul rapides et économiques au possible, i.e. d'optimiser les algorithmes de calcul. On examine le problème d'optimisation avec contraintes données à l'aide des théorèmes mathématiques généraux et on évalue le coût minimal de chaque problème concret appartenant à une classe ou une séquence données. Le fait de se limiter à des cas isolés met en général dans l'impossibilité d'obtenir le résultat cherché. Mais si l'on sait trouver un extrémum lié, i.e. si l'on résout au mieux chaque problème local, on se rapproche par là même du problème global. Cette philosophie de l'optimisation des méthodes numériques, qui est celle de Babuška et Sobolev [XX ; 1], reflète assez fidèlement la question.

Il est parfois impossible de dépasser le cap de l'algorithme voisin de l'optimum. Cela est juste par exemple pour les algorithmes asymptotiquement optimaux. On peut dire que c'est la théorie des estimations asymptotiques qui sert aujourd'hui d'outil puissant dans

l'optimisation algorithmique de diverses classes de problèmes.

La théorie des formules de cubature bâtie par Sobolev [I ; 6] paraît la mieux élaborée du point de vue de l'optimisation. Sobolev ramène l'évaluation des formules de cubature à la minimisation de la fonctionnelle linéaire des erreurs et obtient des estimations d'erreurs dans les formules de cubature relatives à des réseaux réguliers pour les classes de fonctions périodiques indéfiniment dérivables à support borné. L'étude de ces problèmes repose pour une bonne part sur l'évaluation asymptotique des approximations. Signalons les travaux de Bakhvalov [XX ; 2] et Sobol [XVIII ; 10] qui traitent des estimations optimales de la convergence des processus de cubature, des méthodes d'intégration du type Monte-Carlo et des procédés d'intégration numérique optimaux.

On doit à Vinogradov [XX ; 3] suivi par Korobov [XX ; 5] une autre approche des formules de cubature à la lumière de l'Analyse numérique : on construit les formules exactes pour des polynômes trigonométriques finis et on évalue les erreurs sur la classe de fonctions périodiques.

Kolmogorov [XX ; 4] a introduit plusieurs notions générales de la théorie des ensembles pour encadrer le nombre d'opérations arithmétiques exigées par le calcul des solutions. La chose est particulièrement importante dans la recherche dirigée d'algorithmes si la différence de deux bornes tend vers ∞ avec n. S'agissant d'opérateurs différentiels linéaires possédant un inverse compact, Kolmogorov a évalué le nombre d'opérations nécessaires, ce qui permet d'obtenir des algorithmes asymptotiquement voisins de l'optimum en ce qui concerne le critère « nombre d'opérations arithmétiques ».

Bakhvalov [XX ; 2] a étudié un ensemble d'algorithmes de différences finies. Il a en particulier minoré le nombre d'opérations dans le cas de Dirichlet relatif à l'équation de Laplace.

Une direction intéressante de l'optimisation des problèmes de la physique mathématique est liée au nom de Lébédev [IX ; 8] [XVII ; 9] qui prend pour fonctionnelle principale à minimiser le coût de l'algorithme et l'entropie et applique cette méthode à certains problèmes de la théorie de transport.

Dans l'optimisation, on fait souvent abstraction des facteurs tels que l'arrondissement, les particularités des opérations arithmétiques dans les registres de l'ordinateur, etc. Or, ce sont ces facteurs qui déterminent quelquefois l'efficacité de l'algorithme. On se pose donc le problème d'optimiser les calculs mêmes.

La théorie des processus numériques et leur optimisation font l'objet de nombreux travaux dont ceux de Babuška, Sobolev [XX ; 1], Dahlquist [XX], Henrici [III]. Babuška, Vitásek, Práger [III] ont introduit la notion de α_k-suites pour exprimer la relation entre la chaîne de calculs et la précision : quand la première s'allonge, l'autre croît comme une puissance. La théorie des α_k-suites est à l'origine

des notions de stabilité locale et globale qui ont servi de support à l'analyse de tout un éventail d'algorithmes disponibles pour les ordinateurs.

Divers travaux récents (Moore [XX], Nickel [XX]) ont mis à l'ordre du jour les arithmétiques d'intervalles. Il s'agit d'une direction de la théorie de l'estimation du degré d'exactitude d'algorithmes effectifs dont le but principal est de représenter à postériori l'erreur par une paire de bornes.

9.6. Certaines lignes d'évolution du Calcul numérique

Les progrès récents du hardware ont exercé une grande influence sur les domaines numériques de la science qui sont enclins à l'intégration. Le lien entre le matériel de calcul, les méthodes de l'Analyse numérique et des mathématiques appliquées, la théorie de la programmation automatique et des langages devient si étroit que le choix d'une stratégie propre à tel problème est d'un intérêt capital. Bien que l'optimisation des chaînons isolés du calcul reste l'assise profonde de la théorie, on est de plus en plus axé sur l'optimisation globale.

Nul doute que l'optimisation des calculs sur ordinateur constitue un des grands problèmes des disciplines numériques qui stimule la recherche d'algorithmes nouveaux et des mises en œuvre correspondantes.

Une autre tendance est d'aborder toute une classe de problèmes à la fois et de standardiser les algorithmes. Le traitement massif de l'information par les ordinateurs exige un effort de systématisation et d'ordonnancement. La pratique des problèmes de la science et de la technique permet très souvent de s'orienter vers les méthodes universelles valables pour des classes plus ou moins vastes de problèmes de même type mathématique. Il paraît que la stratégie raisonnable est en l'occurrence de générer des algorithmes universels adaptatifs pour des problèmes variés non répétitifs et de mettre au point des algorithmes spéciaux à des problèmes répétitifs. Avec ces deux approches qui se complètent heureusement, on crée un software efficace en utilisant au mieux les ressources de la société.

La standardisation des programmes et des algorithmes, l'implantation progressive des ordinateurs dans divers domaines : problèmes de l'ingénieur, économie, gestion, la puissance accrue du matériel de calcul électronique et un software perfectionné ont permis d'établir des paquets de programmes pour des secteurs entiers de la science et de la technique. De ce côté, il y a lieu d'attendre des résultats importants qui se répercuteront sûrement sur les méthodes numériques.

Les ordinateurs modernes étant rapides et à grande mémoire constituent autant de dépôts d'information de valeur immédiate-

ment accessible, et les systèmes conversationnels fonctionnant en partage de temps permettent de réaliser le dialogue direct homme-machine. Aussi le problème de standardiser le software, en particulier les algorithmes de calcul, est d'actualité brûlante.

Le développement du software pose des problèmes de caractère numérique tels que la construction de réseaux pour des domaines de forme complexe qui recouvrent uniformément dans un sens le domaine de définition de la solution. Si le problème est près d'être résolu en dimension deux, on le met seulement à l'ordre du jour pour la dimension plus grande. L'entrée et la sortie de l'information ont fait penser à la représentation graphique de celle-ci et ont donné naissance à plusieurs procédés d'interpolation sur les classes de fonctions.

Avec la possibilité d'effectuer les transformations analytiques sur ordinateur, on est en fait amené à aborder les problèmes de la physique mathématique par les méthodes bien élaborées de la théorie de l'argument continu. Plus on excelle dans ces transformations, et plus le software s'oriente vers les méthodes en question. Les progrès dans ce domaine ouvrent des champs nouveaux à la création des méthodes numériques puissantes.

Nous voudrions dire en guise de conclusion que le rythme de la recherche en Calcul numérique est fonction de l'état des disciplines mathématiques fondamentales. Or, l'intérêt de ces secteurs de base et leur rythme de développement augmentent considérablement à l'ère du progrès technique. Seuls les efforts conjugués de tous les chercheurs créeront un climat propice aux développements futurs des mathématiques pures et appliquées.

BIBLIOGRAPHIE

I. Analyse fonctionnelle et analyse numérique

[1] КАНТОРОВИЧ Л. В., Функциональный анализ и прикладная математика, *Усп. матем. наук*, 3, 6 (1948).

[2] КЕЛДЫШ М. В., ЛИДСКИЙ В. Б., Вопросы спектральной теории несамосопряженных операторов. In : Труды IV Всесоюзного математического съезда. т. I, М., АН СССР, 1963.

[3] КРЕЙН С. Г., *Линейные дифференциальные уравнения в банаховом пространстве*, М., « Наука », 1967.

[4] НАТАНСОН И. П., *Теория функций вещественной переменной*, М., « Наука », 1974.

[5] НИКОЛЬСКИЙ С. М., *Приближение функций многих переменных и теоремы вложения*, М., « Наука », 1969.

[6] СОБОЛЕВ С. Л., а) *Некоторые применения функционального анализа в математической физике*, Л., Изд-во ЛГУ, 1950.
б) *Введение в теорию кубатурных формул*, М., « Наука », 1974.

[7] ФАГЕ М. К., *Теория линейных операторов*, Новосибирск, 1972.

[8] ANSELONE P. M., *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1967.

[9] BALAKRISHNAN A. V., *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[10] COLLATZ L., *Funktionalanalysis und numerische Mathematik*, Berlin, Springer, 1964.

[11] KANTOROVICH L. V., AKILOV G. P., *Functional analysis in normed spaces*, McMillan, 1964.

[12] KOLMOGOROV A., FOMINE S., *Eléments de la théorie des fonctions et de l'analyse fonctionnelle*, Ed. de Moscou, 1974.

[13] KRASNOSELSKII M. A. et al., *Approximate solution of operator equations*, Walters-Nordhoff-Publishing, Groningen, 1972.

[14] LAVRENTIEV M., CHABAT B., *Méthodes de la théorie des fonctions d'une variable complexe*, Ed. de Moscou, 1972.

[15] LIONS J., *Equations différentielles opérationnelles et problèmes aux limites*, Berlin-Göttingen-Heidelberg, Springer-Verlag, 1961.

[16] LJUSTERNIK L. A., SOBOLEV W. I., *Elemente der Funktionalanalysis*, Berlin, Akad-Verlag, 1968.

[17] MIKHLIN S. G., *Variational Methods in Mathematical Physics*, Pergamon Press, Oxford, 1964.

[18] VARGA R. S., *Functional analysis and approximation theory in numerical analysis*, Soc. ind and appl. math., Philadelphia, 1971.

[19] WEINSTEIN A., STENGER W., *Methods of Intermediate Problems for Eigenvalues : Theory and Ramifications*, Acad. Press, London and New York, 1972.

II. Les équations aux dérivées partielles et la physique mathématique

[1] БИЦАДЗЕ А. В., *Краевые задачи для эллиптических уравнений второго порядка*, М., « Наука », 1966.

[2] ВИШИК М. И., ЛЮСТЕРНИК Л. А., Асимптотическое поведение решений линейных дифференциальных уравнений с большими или быстроменяющимися коэффициентами и граничными условиями, *Усп. матем. наук*, 15, 4 (94), (1960).

[3] КОНДРАТЬЕВ В. А., Краевые задачи для эллиптических уравнений в областях с коническими или угловыми точками, *Тр. Моск. мат. о-ва*, М., 1967, т. 16, 209-292.

[4] ЛАВРЕНТЬЕВ М. А., *Вариационный метод в краевых задачах для систем уравнений эллиптического типа*, М., АН СССР, 1952.

[5] ЛАДЫЖЕНСКАЯ О. А., *Смешанная задача для гиперболического уравнения*, М., Гостехиздат, 1953.

[6] ПЕТРОВСКИЙ И. Г., *Лекции об уравнениях с частными производными*, М., Физматгиз, 1961.

[7] РОЖДЕСТВЕНСКИЙ Б. Л., ЯНЕНКО Н. Н., *Системы квазилинейных уравнений*, М., « Наука », 1968.

[8] СМИРНОВ В. И., *Курс высшей математики*, т. 5, М., Гостехиздат, 1948.

[9] СОБОЛЕВ С. Л., *Уравнения математической физики*, М., « Наука », 1966.

[10] ТИХОНОВ А. Н., САМАРСКИЙ А. А., *Уравнения математической физики*, М., « Наука », 1966.

[11] ФИХТЕНГОЛЬЦ Г. М., *Курс дифференциального и интегрального исчисления*, т. 3, М., Физматгиз, 1963.

[12] BOSSAVIT A., Régularisation d'équations variationnelles et applications, Centre National de la Recherche Scientifique, Institut Blaise Pascal, Juin 1970.

[13] COURANT R., *Partial differential equations*, J. Wiley, N.Y.-London, 1962.

[14] COURANT R., HILBERT D., *Methods of mathematical physics*, vol. 1, New York-London, Interscience publ., 1953.

[15] FRIEDRICHS K., Non-linear hyperbolic differential equations for functions of two independent variables, *Amer. J. Math.*, **70** (1948).

[16] GODOUNOV S., *Equations de la physique mathématique*, Ed. de Moscou, 1973.

[17] KANTOROVICH L. V., KRYLOV V. I., *Approximate methods of higher analysis*, Groningen (Nordhoff), 1958.

[18] LAVRENTIEV M. M., Some improperly posed problems of Mathematical Physics, Springer, Tracts in Natural Philosophy, 2, (1967).

[19] LIONS J., *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Paris, Dunod, 1969.

[20] LIONS J., MAGENES E., *Problèmes aux limites non homogènes et applications*, t. 1, Paris, Dunod, 1968.

[21] MIGNOT A., Méthodes d'approximation des solutions de certains problèmes aux limites linéaires, *Rend. semin. mat. Univ. Padova*, XL, (1968).

[22] MILLER J. H. ed., Topics in Numerical Analysis. Proc. Roy. Irish Acad. Conf. on Numer. Analysis 1972, Acad. Press, London and New York, 1973.

[23] MILLER J. H. ed., Topics in Numerical Analysis. II. Proc. Roy. Irish Acad. Conf. on Numer. Analysis 1974, Acad. Press, London and New York, 1975.

[24] SMIRNOV V., Cours de matématiques supérieures, t. 1-4, Ed. de Moscou, 1970-1975.

[25] VEKUA I. N., *New Methods for solving elliptic Equations*, John Wiley, New York, 1967.

[26] VLADIMIROV V., *Distributions en physique mathématique*, Ed. de Moscou, 1979.
[27] WLADIMIROV W. S., *Gleichungen der mathematischen Physik*, Berlin, 1972.

III. Méthodes numériques (monographies et manuels)

[1] ВАЛИУЛЛИН А. Н., *Схемы повышенной точности для задач математической физики*, Новосибирск, 1973.
[2] ВОРОБЬЕВ Ю. В., *Метод моментов в прикладной математике*, М., Физматгиз, 1958.
[3] ГОДУНОВ С. К., *Разностные методы решения уравнений газовой динамики*, Новосибирск, Изд-во НГУ, 1962.
[4] ГОДУНОВ С. К., ЗАБРОДИН А. В., ИВАНОВ М. Я. et al., *Численное решение многомерных задач газовой динамики*, М., « Наука », 1976.
[5] ДОРОДНИЦЫН А. А., а) Об одном методе численного решения некоторых нелинейных задач аэрогидродинамики. In : Труды III Всесоюзного математического съезда, т. II, М., Изд-во АН СССР, 1956.
б) *Лекции по численным методам решения уравнений вязкой жидкости*, М., ВЦ АН СССР, 1969.
[6] ДЬЯКОНОВ Е. Г., а) *Итерационные методы решения разностных аналогов краевых задач для уравнений эллиптического типа*, Киев, Ин-т кибернетики АН УССР, 1970.
б) Разностные методы решения краевых задач. Вып. 1 (стационарные задачи), М., МГУ, 1971, вып. 2 (нестационарные задачи), М., МГУ, 1972.
[7] ИЛЬИН В. П., а) *Разностные методы решения эллиптических уравнений*, Новосибирск, Изд-во НГУ, 1970.
б) *Численные методы решения задач электрооптики*, Новосибирск, « Наука », 1974.
[8] КОНОВАЛОВ А. Н., *Численное решение задач упругости*, Новосибирск, Изд-во НГУ, 1968.
[9] МАРЧУК Г. И., а) *Методы расчета ядерных реакторов*, М., Атомиздат, 1961.
б) *Методы вычислительной математики*, Новосибирск, « Наука », 1973.
[10] Некоторые проблемы вычислительной и прикладной математики. Sous la direction de M. M. ЛАВРЕНТЬЕВ, Новосибирск, « Наука », 1975.
[11] ПОЛОЖИЙ Г. Н., *Численное решение двумерных и трехмерных краевых задач математической физики и функций дискретного аргумента*, Киев, Изд-во КГУ, 1962.
[12] САМАРСКИЙ А. А., *Введение в теорию разностных схем*, М., « Наука », 1971.
[13] САМАРСКИЙ А. А., ГУЛИН А. В., *Устойчивость разностных схем*, М., « Наука », 1973.
[14] САМАРСКИЙ А. А., ПОПОВ Ю. П., *Разностные схемы газовой динамики*, М., « Наука », 1975.
[15] САУЛЬЕВ В. К., *Интегрирование уравнений параболического типа методом сеток*, М., Физматгиз, 1960.
[16] ЧЕРНОУСЬКО Ф. Л., БАНИЧУК Н. В., *Вариационные задачи механики и управления*, М., « Наука », 1973.
[17] ЯНЕНКО Н. Н.,Введение в разностные методы математической физики, ч. 1, 2, Новосибирск, Изд. Новосиб. ун-та, 1968.
[18] AHLBERG J. H., NILSON E. N., WALSH J. L., a) Extremal orthogonalities, *J. Math. Anal. Appl.*, **12**, 1 (1965).
b) *The theory of splines and their applications*, Academic Press, New York-London, 1967.
[19] AUBIN J. P., *Approximation of Elliptic Boundary-Value Problems*, Wiley-Interscience, 1972.

[20] BABUŠKA I., VITÁSEK E., PRÁGER M., *Numerical processus in differential Equations*, Intersc. Publ., 1966.
[21] BAKHVALOV N., *Méthodes numériques*, Ed. de Moscou, 1976.
[22] BALAKRISHNAN A. V., NEUSTADT L. W., *Computing methods in optimization problems*, Acad. Press, 1964.
[23] BELLMAN R., KALABA R., LOCKETT J., *Numerical inversion of the Laplace Transform*, Am. Elsevier Publ. Co. Inc., N.Y., 1966.
[24] BENSOUSSAN A., LIONS J. L., TEMAM R., Sur les méthodes de décomposition, de décentralisation et de coordination et applications. Sur les méthodes numériques en sciences physiques et économiques, Paris, Dunod, 1974.
[25] BEREZIN I. S., ZHIDKOV N. P., *Computing methods*, Addison-Wesley Publ. Co., Palo Alto, 1965.
[26] BLUM E. K., *Numerical Analysis and Computation : Theory and Practice*, Addison-Wesley Publ. Co. Inc., London, 1972.
[27] BRAMBLE J. H. ed., Numerical Solution of Partial Differential Equations. Proc. of a Symp. Held at the Univ. of Maryland, Acad. Press, London-New York, 1966.
[28] CÉA J., *Optimisation. Théorie et algorithmes*, Dunod, Paris, 1971.
[29] COLLATZ L., *Numerische Behandlung von Differentialgleichungen*, Berlin, Springer, 1951.
[30] DAHLQUIST G., BJÖRCK A., *Numerical Methods*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1974.
[31] FORSYTHE G. E., WASOW W. R., *Finite Difference Methods for partial differential equations*, N.Y.-London, 1960.
[32] FOX L., MAYERS D. F., *Computing Methods for Scientists and Engineers*, Oxford, 1968.
[33] GODOUNOV S., RIABENKI V., *Schémas aux différences*, Ed. de Moscou, 1977.
[34] GODUNOV S. K., RYABENKI V. S., *Introduction to the theory of difference schemes*, Interscience, New York, 1964.
[35] GUELFOND A., *Calcul des différences finies*, Paris, Dunod, 1962.
[36] HENRICI P., *Error propagation for difference methods*, J. Wiley and Sons, New York, 1963.
[37] HOUSEHOLDER A. S., *Principles of numerical analysis*, McGraw-Hill, New York, 1953.
[38] ISAACSON E., KELLER H. B., *Analysis of numerical methods*, Wiley, N.Y., 1966.
[39] KANTOROVICH L. V., KRYLOV V. I., *Approximate methods of higher analysis*, Groningen (Nordhoff), 1958.
[40] KELLER H. B., *Numerical Methods for Two-Point Boundary-Value Problems*, Blaisdell Publ. Comp., New York, 1968.
[41] LANCZOS C., *Applied Analysis*, Prentice Hall, Englewood Cliffs, N.Y., 1956.
[42] MARCHOUK G. I., KUZNETSOV Yu. A., *Méthodes itératives et fonctionnelles quadratiques. Sur les méthodes numériques en sciences physiques et économiques*, Paris, Dunod, 1974.
[43] MARCHUK G. I., a) *Méthodes numériques pour la prévision du temps*, A. Colin, Paris, 1970.
b) Methods and problems of computational mathematics. Actes du congrès international des mathématiciens. Septembre 1970. Nice, t. 2, Gauthier-Villars, Paris, 1971.
[44] MILLER J., STRANG G., Matrix theorems for partial differential and difference equations, *Math. scand.*, 18, 2 (1966).
[45] MITCHELL A. R., *Computational methods in partial differential equations*, Wiley, London, 1970.

[46] MYSOVSKIH I. P., *Lectures on numerical methods*, Groningen, Wolhers-Nordhoff, 1969.
[47] PCHÉNITCHNY B., DANILINE Y., *Méthodes numériques dans les problèmes d'extrémum*, Ed. de Moscou, 1977.
[48] POLAK E., *Computational methods in optimization : a unifed approach*, Acad. Press, New York, 1971.
[49] PONTRIAGUINE L., BOLTIANSKI V., GAMKRÉLIDZÉ R., MICHTCHENKO E., *Théorie mathématique des processus optimaux*, Ed. de Moscou, 1974.
[50] RICHTMYER R. D., *Difference methods for initial value problems*, N.Y., 1957.
[51] RICHTMYER R., MORTON K. W., *Difference methods for initial-value problems*, Wiley, N.Y., 1967.
[52] SAMARSKI A., ANDRÉEV V., *Méthodes aux différences pour équations elliptiques*, Ed. de Moscou, 1978.
[53] STRANG G., FIX G. J., *An Analysis of the Finite Element Methods*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1973.
[54] TRAUB J. F., *Iterative Methods for the Solution of Equations*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1964.
[55] VARGA R. S., *Matrix iterative analysis*, Englewood Cliffs, N.Y., 1962.
[56] YANENKO N. N., *Méthode à pas fractionnaires*, Armand Colin, Paris, 1968.

IV. Méthode des différences finies

[1] БЕЛОЦЕРКОВСКИЙ О. М., ЧУШКИН П. И., Численный метод интегральных соотношений, *Ж. вычисл. матем. и матем. физ.*, 2, 5 (1962).
[2] ВАЛИУЛЛИН А. Н., ЯНЕНКО Н. Н., Экономичные разностные схемы повышенной точности для полигармонического уравнения, *Изв. Сиб. отд. АН СССР. Сер. техн. наук*, 13, 3 (1967).
[3] ВАЛИЦКИЙ Ю. Н., О сходимости разностных аппроксимаций собственных значений и собственных функций двумерного эллиптического оператора, *Докл. АН СССР*, 198, 2 (1971).
[4] ВОЛКОВ Е. А., а) Исследование одного способа повышения точности метода сеток при решении уравнения Пуассона, *Вычислительная математика*, 1957, т. 1.
б) Решение задачи Дирихле методом уточнений разностями высших порядков, ч. I, II, *Дифференц. уравнения*, 1, 7, 8 (1965).
в) Метод неравномерных сеток для конечных и бесконечных областей с коническими точками, *Дифференц. уравнения*, 10, 2 (1966).
г) Развитие метода сеток для уравнения Лапласа на конечных и бесконечных областях с кусочно-гладкой границей. (Thèse doct.), М., 1967.
[5] ГОДУНОВ С. К., Разностные методы численного решения задач газовой динамики, *Ж. вычисл. матем. и матем. физ.*, 2, 7 (1962).
[6] ГОДУНОВ С. К., ЗАБРОДИН А. В., О разностных схемах второго порядка точности для многомерных задач, *Ж. вычисл. матем. и матем. физ.*, 2, 4 (1962).
[7] ГОДУНОВ С. К., ПРОКОПОВ Г. П., О расчетах конформных отображений и построении разностных сеток, *Ж. вычисл. матем. и матем. физ.*, 7, 5 (1967).
[8] ДЕМЬЯНОВИЧ Ю. К., Метод сеток для некоторых задач математической физики, *Докл. АН СССР*, 159, 2 (1964).
[9] КОНОВАЛОВ А. Н., Метод фиктивных областей в задачах кручения In : Численные методы механики сплошной среды, 4, 2 (1973).
[10] КОПЧЕНОВ В. Д., Приближение решения задачи Дирихле методом фиктивных областей, *Дифференц. уравнения*, 4, 1 (1968).

[11] КУЗНЕЦОВ Ю. А., МАЦОКИН А. М., а) Решение уравнения Гельмгольца методом фиктивных областей. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1972.
б) Об оптимизации метода фиктивных компонент и его применения. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1977.

[12] КУЗНЕЦОВ Ю. А., ШАЙДУРОВ В. В., О равномерной сходимости разностных схем. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1972.

[13] КУРОПАТЕНКО В. Ф., Метод построения разностных схем для численного интегрирования уравнений газодинамики, *Изв. высш. учебных заведений. Математика*, **3**, 28, (1962).

[14] ЛАНДАУ Л. Д., МЕЙМАН Н. Н., ХАЛАТНИКОВ И. М., Численные методы интегрирования уравнений в частных производных сеток. In : Труды III Всесоюзного математического съезда, т. II, М., АН СССР, 1956.

[15] ЛЕБЕДЕВ В. И., а) Метод сеток для уравнений типа С. Л. Соболева, *Докл. АН СССР*, **114**, 6 (1957).
б) О методе сеток для одной системы уравнений в частных производных, *Изв. АН СССР. Сер. матем.*, **22**, 5 (1958).
в) О задаче Дирихле и Неймана на треугольных и шестиугольных сетках, *Докл. АН СССР*, **138**, 1 (1961).
г) Разностные аналоги ортогональных разложений, основных дифференциальных операторов и некоторых краевых задач математической физики, *Ж. вычисл. матем. и матем. физ.*, **4**, 3, 4 (1964).

[16] ЛЮСТЕРНИК Л. А., О разностных аппроксимациях операторов Лапласа, *Усп. матем. наук*, **IX**, 2 (1954).

[17] МАРЧУК Г. И., ДЫМНИКОВ В. П., ГАЛИН В. Я. et al., *Гидродинамическая модель общей циркуляции атмосферы и океана*, Новосибирск, 1975.

[18] МАРЧУК Г. И., РИВИН Г. С., ЮДИН М. И., Численные эксперименты с балансными схемами, *Изв. АН СССР. Сер. ФАО*, **11** (1973).

[19] МАРЧУК Г. И., ШАЙДУРОВ В. В., О численном решении эволюционной задачи с ограниченным оператором, *Докл. АН СССР*, **216**, 1 (1974).

[20] МАЦОКИН А. М., а) К развитию метода фиктивных областей. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1975.
б) Вариационно-разностный метод решения эллиптических уравнений в трехмерных областях. In : Вариационно-разностные методы в математической физике, Новосибирск, ВЦ СО АН СССР, 1976.
в) О построении и методах решения систем вариационно-разностных уравнений. (Thèse cand.), Новосибирск, ВЦ СО АН СССР, 1975.

[21] ПЕНЕНКО В. В., Вычислительные аспекты в задачах математического моделирования динамики атмосферных процессов. (Thèse doct.), Новосибирск, 1976.

[22] РИВКИНД В. Я., а) Приближенный метод решения задачи Дирихле и об оценках скорости сходимости решений разностных уравнений к решениям эллиптических уравнений с разрывными коэффициентами, *Вестник Ленингр. ун-та. Сер. матем.*, 3 (1964).
б) Об оценке скорости сходимости однородных разностных схем для эллиптических и параболических уравнений с разрывными коэффициентами. In : Проблемы математического анализа, Л., ЛГУ, 1966.

[23] РУХОВЕЦ Л. А., Замечание к методу фиктивных областей, *Дифференц. уравнения*, **3**, 4 (1967).

[24] САМАРСКИЙ А. А., а) О монотонных разностных схемах для эллиптических и параболических уравнений в случае несамосопряженного эллиптического оператора, *Ж. вычисл. матем. и матем. физ.*, **5**, 3 (1965).

б) О точности метода сеток для задачи Дирихле в произвольной области, *Apl. Mat.*, **10**, 3 (1965).
в) Некоторые вопросы теории разностных схем, *Ж. вычисл. матем. и матем. физ.*, 6, 4 (1966).

[25] САУЛЬЕВ В. К., а) Об одном методе автоматизации решения краевых задач на быстродействующих вычислительных машинах, *Докл. АН СССР*, **142**, 3 (1962).
б) О решении некоторых краевых задач на быстродействующих вычислительных машинах методом фиктивных областей, *Сиб. мат. ж.*, 4, 4 (1963).

[26] СИНЯЕВ В. Н., Об одном принципе построения конечно-разностных схем, основанных на законах сохранения полной энергии. In : Численные методы механики сплошной среды, Новосибирск, ВЦ СО АН СССР, т. 5, № 2 (1974).

[27] ТИХОНОВ А. Н., САМАРСКИЙ А. А., а) О разностных схемах для уравнений с разрывными коэффициентами, *Докл. АН СССР*, **108**, 3 (1956).
б) Однородные разностные схемы на неравномерных сетках, *Ж. вычисл. матем. и матем. физ.*, 2, 5 (1962).

[28] УРВАНЦЕВ А. Л., ШАЙДУРОВ В. В., Уточнение приближенного решения квазилинейного уравнения Пуассона. In : Вариационно-разностные методы в математической физике, Новосибирск, ВЦ СО АН СССР, 1976.

[29] ЧУДОВ Л. А., КУДРЯВЦЕВ В. П., Об ошибках округления при решении разностными методами задач с начальными условиями для эллиптических уравнений и систем. In : Численные методы в газовой динамике, М., Изд. Моск. ун-та, 1963.

[30] ШАЙДУРОВ В. В., Об одном методе повышения точности разностных решений. In : Численные методы механики сплошной среды, 3, 2 (1972).

[31] ЯНЕНКО Н. Н., СУЧКОВ В. А., ПОГОДИН Ю. Я., О разностном решении уравнения теплопроводности в криволинейных координатах, *Докл. АН СССР*, **128**, 5 (1959).

[32] BRYAN K., A scheme for numerical integration of the equations of motion on an irregular grid free of non-linear instability, *Monthly Weather Review*, 94, 1 (1966).

[33] CROWLEY W., Second order numerical advection, *J. Comp. Phys.*, **1**, 4 (1967).

[34] FOX L., HENRICI P., MOLER C., Approximations and bounds of eigenvalues of elliptic operators, *SIAM J. Numer. Anal.*, 4, 1 (1967).

[35] FROMM J. E., Numerical method for computing non-linear time dependent buoyant circulation of air in rooms, *JBM J. of Research and Development*, **15**, 5 (1971).

[36] JOICE D. C., Survey of extrapolation processes in numerical analysis, *SIAM Review*, **13**, 4 (1971).

[37] KELLER H., A new difference scheme for parabolic problems. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.

[38] KELLOGG R., Singularities in interface problems. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.

[39] KURIHARA Y., HOLLOWAY I., Numerical integration of a nine-level global primitive equations model formulated by the box method, *Monthly Weather Review*, 95, 8 (1967).

[40] MARCHUK G. I., SHAYDOUROV V. V., Increasing of the accuracy of the projective-difference schemes, *Lecture Notes in Computer Science*, **11**, Springer Verlag, 1974.

[41] MIGNOT A., Méthodes d'approximation des solutions de certains problè-

mes aux limites linéaires, *Rend. semin. mat. Univ. Padova*, XL (1968).
[42] RAVIART P. A., Sur l'approximation de certaines équations d'évolution linéaires et non linéaires, *J. de Mathem. Pures et Appl.*, 46, 1 (1967).
[43] RICHARDSON L. F., The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a mansory dam, *Philos. Trans. Roy. Soc.*, London, Ser. A, 210 (1910).
[44] WACHSPRESS E. L., The Numerical Solution of Boundary Value Problems. Mathematical Methods for Digital Computers, N.Y., 1960.

V. Méthodes variationnelles aux différences

[1] АГОШКОВ В. И., О вариационной форме интегрального тождества Г. И. Марчука, préprint ВЦ СО АН СССР, Новосибирск, 1977.
[2] ДЬЯКОНОВ Е. Г., Некоторые классы операторов, эквивалентных по спектру, и их применения. In : Вариационно-разностные методы в математической физике, Новосибирск, ВЦ СО АН СССР, 1976.
[3] КЕЛДЫШ М. В., О методе Галеркина для решения краевых задач, *Изв. АН СССР. Сер. матем.*, 6 (1942).
[4] ЛЕБЕДЕВ В. И., Разностные аналоги ортогональных разложений, основных дифференциальных операторов и некоторых краевых задач математической физики, *Ж. вычисл. матем. и матем. физ.*, 4, 3, 4 (1964).
[5] МАРЧУК Г. И., АГОШКОВ В. И., О выборе координатных функций в обобщенном методе Бубнова-Галеркина, *Докл. АН СССР*, 232, 6 (1977).
[6] МАУОКИН А. М., а) Автоматизация триангуляции областей с гладкой границей при решении уравнений эллиптического типа, préprint n° 15 du séminaire *Вычислительные методы прикладной математики*, ВЦ СО АН СССР, 1975.
б) Вариационно-разностный метод решения эллиптических уравнений в круге, préprint n° 13 du séminaire *Вычислительные методы прикладной математики*, Новосибирск, ВЦ СО АН СССР, 1975.
[7] ОГАНЕСЯН Л. А., а) Численный расчет плит. In : Решение инженерных задач на электронно-вычислительных машинах, Л., 1963.
б) Вариационно-разностная схема на регулярной сетке для задачи Дирихле, *Ж. вычисл. матем. и матем. физ.*, 11, 6 (1971).
[8] ОГАНЕСЯН Л. А., РИВКИНД В. Я., РУХОВЕЦ Л. А., Вариационно-разностные методы решения эллиптических уравнений, I, II. In : Дифференциальные уравнения и их применение, вып. 5, 8, Вильнюс, 1974.
[9] ОГАНЕСЯН Л. А., РУХОВЕЦ Л. А., а) О вариационно-разностных схемах для линейных эллиптических уравнений второго порядка в двумерной области с кусочно-гладкой границей, *Ж. вычисл. матем. и матем. физ.*, 8, 1 (1968).
б) Исследование скорости сходимости вариационно-разностных схем для эллиптических уравнений второго порядка в двумерной области с гладкой границей, *Ж. вычисл. матем. и матем. физ.*, 9, 5 (1969).
[10] РУХОВЕЦ Л. А., Исследование скорости сходимости вариационно-разностных схем для двумерных эллиптических уравнений второго порядка. (Thèse cand.), Л., 1970.
[11] СМЕЛОВ В. В., Аппроксимация кусочно-гладких функций тригонометрическими многочленами и использование последних в вариационных методах (Préprint), Новосибирск, 1975.
[12] ШАЙДУРОВ В. В., Экстраполяция Ричардсона для проекционно-разностной задачи Штурма-Лиувилля. In : Вариационно-разностные методы в математической физике, Новосибирск, ВЦ СО АН СССР, 1974.
[13] AUBIN J. P., a) Approximation des espaces de distributions et des opérateurs différentiels, *Bull. Soc. Math. France*, Mémoire, 12 (1967).

b) Behavior of the error of the approximate solutions of boundary value problems for linear elliptic equations by Galerkin's and finite difference methods, *Ann. Scuola. Norm. Super.*, Pisa, **21**, 4 (1967).
c) Best approximation of linear operators in Hilbert space, *SIAM J. Numer. Anal.*, 5, 3 (1968).

[14] AUBIN J. P., BURCHARD H. G., Some aspects of the method of the hypercircle applied to elliptic variational problems. Numerical solution of partial differential equations — II. SYNSPADE 1970, Acad. Press, New York-London, 1971.

[15] BABUŠKA I., a) The finite element method for elliptic differential equations. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.
b) The rate of convergence for finite element method, *SIAM J. Numer. Anal.*, 8, 2 (1971).

[16] BIRKHOFF G., SCHULTZ M. H., VARGA R. S., Hermite interpolation in one and two variables with applications to partial differential equations, *Numer. Math.*, **11**, 3 (1968).

[17] BRAMBLE J., A second order finite difference analog of the first biharmonic boundary value problems, *Numer. Math.*, **9**, 3 (1966).

[18] BRAMBLE J., HUBBARD B., On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation, *Numer. Math.*, **4**, 4 (1962).

[19] BRAMBLE J., SCHATZ A., On the numerical solution of elliptic boundary value problems by least squares approximation of the data. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.

[20] BRÉZIS H., STAMPACCHIA G., Sur la régularité de la solution d'inéquations elliptiques, *Bull. Soc. Math. France*, 96 (1968).

[21] CÉA J., Approximation variationnelle des problèmes aux limites, *Ann. Inst. Fourier*, Grenoble, **14**, 2 (1964).

[22] CÉA J., GLOWINSKI R., a) Sur des méthodes d'optimisation par relaxation, *Revue Française d'Automatique, Informatique et Recherche Opérationnelle*, 1973, R-3.
b) Méthodes numériques pour l'écoulement laminaire d'un fluide rigide viscoplastique incompressible, *Int. J. of Comp. Math.*, Ser. B, 3, (1974).

[23] COURANT R., Variational methods for the solution of problems of equilibrium and vibrations, *Bull. Amer. Math. Soc.*, **49** (1943).

[24] DOUGLAS J., DUPONT T., Alternating-direction Galerkin methods on rectangles. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.

[25] DUVAUT G., LIONS J. L., *Les inéquations en mécanique et en physique*, Paris, Dunod, 1972.

[26] GLOWINSKI R., Introduction to the Approximation of Elliptic Variational Inequalities. Report 76006, Laboratoire d'Analyse Numérique de l'Université de Paris, 6 (1976).

[27] GLOWINSKI R., LIONS J. L., TRÉMOLIÈRES R., *Analyse numérique des inéquations variationnelles*, t. 1 et 2, Dunod, Paris, 1976.

[28] GLOWINSKI R., MARROCO A., Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires, *Revue Française d'Automatique, Informatique et Recherche Opérationnelle*, 1975, R-2.

[29] HUBBARD B., Remarks on the convergence in the discrete Dirichlet problem. Numerical solution of partial differential equations. Edited by James H. Bramble, Academic Press, New York-London, 1965.

[30] LIONS J. L., *Contrôle optimal de systèmes commandés par des équations aux dérivées partielles*, Paris, Dunod, 1969.

[31] LIONS J. L., STAMPACCHIA G., Variational Inequalities, *Comm. Pure Applied Math.*, **XX** (1967).
[32] LIONS J. L., TEMAM R., Une méthode d'éclatement des opérateurs et des contraintes en calcul des variations, *C.R. Acad. Sci. Paris*, **263**, 17 (1966).
[33] SMIRNOV V., *Cours de mathématiques supérieures*, v. IV, première partie, Ed. de Moscou, 1975.
[34] STAMPACCHIA G., Formes bilinéaires coercitives sur les ensembles convexes, *C.R. Acad. Sci. Paris*, **258** (1964).
[35] STRANG G., The finite element method and approximation theory. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.
[36] STRANG G., FIX G., A Fourier analysis of the finite element variational method, preprint (1970).
[37] ZIENKIEWICZ O. C., The finite element method in engineering science, McGraw-Hill, London, 1971.
[38] ZLAMAL M., a) On the finite element method, *Numer. Math.*, **12**, 5 (1968).
b) On some finite element procedures for solving second order boundary value problems, *Numer. Math.*, **14**, 1 (1969).

VI. Stabilité des schémas aux différences

[1] ИЛЬИН А. М., Устойчивость разностных схем задачи Коши для систем дифференциальных уравнений в частных производных, *Докл. АН СССР*, **164**, 3 (1965).
[2] РИХТМАЙЕР Р. Д., О нелинейной неустойчивости разностных схем. In : Некоторые вопросы вычислительной и прикладной математики, Новосибирск, « Наука », 1966.
[3] САМАРСКИЙ А. А., Необходимые и достаточные условия устойчивости двухслойных разностных схем, *Докл. АН СССР*, **181**, 4 (1968).
[4] СЕРДЮКОВА С. И., Исследование устойчивости в C явных разностных схем с постоянными действительными коэффициентами, устойчивых в l_2, *Ж. вычисл. матем. и матем. физ.*, 3, 2 (1963).
[5] ФЕДОРЮК М. В., Об устойчивости в C задачи Коши для разностных уравнений и уравнений с частными производными, *Ж. вычисл. матем. и матем. физ.*, **7**, 3 (1967).
[6] ФИЛИППОВ А. Ф., Об устойчивости разностных уравнений, *Докл. АН СССР*, **100**, 6 (1955).
[7] KELLER H. B., THOMÉE V., Unconditionally stable difference method for mixed problems for quasilinear hyperbolic systems in two dimensions, *Comm. Pure Appl. Math.*, **15**, 1 (1962).
[8] KREISS H. O., a) Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren, *Nordisk Tidskr. Informations Behandlung*, **2**, 2 (1962).
b) On difference approximation of the dissipative type of hyperbolic differential equations, *Comm. Pure Appl. Math.*, **17**, 3 (1964).
c) Initial boundary value problems for partial differential and difference equations in one space dimension. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.
[9] LAX P. D., On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients, *Comm. Pure Appl. Math.*, **14**, 3 (1961).
[10] LAX P. D., NIRENBERG L., On stability for difference schemes, a sharp form of Gårding's inequality, *Comm. Pure Appl. Math.*, 19, 41 (1966).
[11] LAX P. D., WENDROFF B., On the stability of difference schemes with variable coefficients, *Comm. Pure Appl. Math.*, **15**, 4 (1962).

[12] RJABENKI V. S., FILIPPOW A. F., *Über die Stabilität von Differenzengleichungen*, Berlin, 1960.
[13] STRANG G., Difference methods for mixed boundary value problem, *Duke Math. J.*, 27, 2 (1960).
[14] THOMÉE V., Generally unconditionally stable difference operators, *SIAM J. Numer. Anal.*, 4, 1 (1967).

VII. Stabilité et convergence

[1] АНДРЕЕВ В. Б., О сходимости разностных схем, аппроксимирующих вторую и третью краевые задачи для эллиптических уравнений, *Ж. вычисл. матем. и матем. физ.*, 8, 6 (1968).
[2] ГОДУНОВ С. К., РЯБЕНЬКИЙ В. С., а) Канонические виды систем линейных обыкновенных разностных уравнений с постоянными коэффициентами, *Ж. вычисл. матем. и матем. физ.*, 3, 2 (1963).
б) Спектральные признаки устойчивости краевых задач для несамосопряженных разностных уравнений, *Усп. матем. наук*, XVIII, 3 (111) (1963).
[3] ЛАДЫЖЕНСКАЯ О. А., Метод конечных разностей в теории уравнений с частными производными, *Усп. матем. наук*, XII, 5 (1957).
[4] РЯБЕНЬКИЙ В. С., а) Структура спектров семейств несамосопряженных разностных операторов. In : Материалы к совместному советско-американскому симпозиуму по уравнениям с частными производными, Новосибирск, 1963.
б) Спектр семейства разностных операторов над функциями на сеточном графе, *Ж. вычисл. матем. и матем. физ.*, 7, 6 (1967).
[5] САМАРСКИЙ А. А., Некоторые вопросы общей теории разностных схем. In : Дифференциальные уравнения с частными производными (труды симпозиума, посвященного 60-летию академика С. Л. Соболева), М., « Наука », 1970.
[6] СОБОЛЕВ С. Л., Некоторые замечания о численном решении интегральных уравнений, *Изв. АН СССР. Сер. матем.*, 20, 4 (1956).
[7] ЯНЕНКО Н. Н., БОЯРИНЦЕВ Ю. Е., О сходимости разностных схем для уравнения теплопроводности с переменными коэффициентами, *Докл. АН СССР*, 139, 6 (1961).
[8] ЯНЕНКО Н. Н., ШОКИН Ю. И., а) О связи корректности первых дифференциальных приближений и устойчивости разностных схем для гиперболических систем уравнений, *Математические заметки*, 4, 5 (1968).
б) О корректности первых дифференциальных приближений разностных схем, *Докл. АН СССР*, 182, 4 (1968).
[9] COURANT R., FRIEDRICHS K., LEWY H., Über die partiellen Differenzengleichungen der mathematischen Physik, *Math. Ann.*, 100, 32 (1928).
[10] DU FORT E. C., FRANKEL S. P., Stability conditions in the numerical treatment of parabolic differential equations, *Math. Tables and Other Aids Comput.*, 7, 43 (1953).
[11] JOHN F., On the integration of parabolic equations by difference methods. I. Linear and quasilinear equations for the infinite interval, *Comm. Pure Appl. Math.*, 5, 2 (1952).
[12] LAX P. D., RICHTMYER R. D., Survey of the stability of linear finite difference equations, *Comm. Pure Appl. Math.*, 9, 2 (1956).
[13] LAX P. D., WENDROFF B., System of conservations laws, *Comm. Pure. Appl. Math.*, 13, 2 (1960).
[14] LEES M., a) A priori estimate for the solution of difference approximations to parabolic partial differential equations, *Duke Math. J.*, 27, 3 (1960).

b) Energy inequalities for the solution of differential equations, *Trans. Amer. Math. Soc.*, **94**, 1 (1960).
[15] LIONS J., Equations différentielles opérationnelles dans les espaces de Hilbert, Centro Int. Mat. Estivo, Varenna (1963). (Equazioni differenziali astratte. Cremonese, Roma, 1963.)
[16] NEUMANN J., RICHTMYER R. D., A method for the numerical calculation of hydrodynamic shocks, *J. Appl. Phys.*, **21**, 3 (1950).
[17] O'BRIEN G. G., HYMAN M. A., KAPLAN S., A study of the numerical solution of partial differential equations, *J. Math. and Phys.*, **29**, 4 (1951).
[18] PHILLIPS N. A., The atmosphere and the sea in motion. Scientific Contributions to the Rossby Memorial Volume, The Rockfeller Institute, 1959.
[19] STRANG G., a) Accurate partial difference methods. I. Linear Cauchy problems, *Arch. Ration. Mech. Anal.*, **12**, 5 (1963).
b) Implicit difference methods for initial-boundary value problems, *J. Math. Anal. Appl.*, **16**, 1 (1966).
[20] THOMÉE V., On the rate of convergence of difference schemes for hyperbolic equations. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.
[21] WENDROFF B., Well-posed and stable difference operators, *SIAM J. Numer. Anal.*, 5, 1 (1968).
[22] WIDLUND O. B., Stability of parabolic difference schemes in the maximum norm, *Numer. Math.*, 8, 2 (1968).

VIII. Méthodes numériques d'algèbre linéaire

[1] АБРАМОВ А. А., Идеи теории возмущений в некоторых алгоритмах линейной алгебры. In : Вычислительные методы линейной алгебры, вып. 1, М., ВЦ АН СССР, 1968.
[2] БАХВАЛОВ Н. С., Основы вычислительной математики. Курс лекций, М., МГУ, 1970.
[3] ВОЕВОДИН В. В., а) Ошибки округления и устойчивость в прямых методах линейной алгебры, М., МГУ, 1969.
б) *Численные методы алгебры. Теория и алгоритмы*, М., « Наука », 1966.
[4] ДОРОДНИЦЫН А. А., К задаче вычисления собственных чисел и собственных векторов матриц, *Докл. АН СССР*, **126**, 6 (1959).
[5] ИКРАМОВ Х. Д., Матричные нормы и методы типа Якоби, М., МГУ, 1969.
[6] ИЛЬИН В. П., О некоторых оценках для методов сопряженных градиентов, *Ж. вычисл. матем. и матем. физ.*, **16**, 4 (1976).
[7] КИМ Г., О распределении ошибок округления итерационных методов решения систем линейных алгебраических уравнений, М., МГУ, 1969.
[8] КУБЛАНОВСКАЯ В. Н., Применение ортогональных преобразований для решения задач алгебры. (Thèse doct.), Л., 1972.
[9] КУЗНЕЦОВ Ю. А., *Итерационные методы решения несовместных систем линейных уравнений. Некоторые проблемы вычислительной и прикладной математики*, Новосибирск, « Наука », 1975.
[10] КУЗНЕЦОВ Ю. А., МАРЧУК Г. И., Итерационные методы решения систем линейных уравнений с особенными матрицами, *Acta Universitatis Carolinae. Mathematica et Physica*, Praha, **15** (1974).
[11] МАРЕК И., Итерации линейных ограниченных операторов и процесс Келлога. (Thèse), Прага, 1962.
[12] МАРЧУК Г. И., КУЗНЕЦОВ Ю. А., Итерационные методы и квадратичные функционалы. In : Методы вычислительной математики, Новосибирск, « Наука », 1975.
[13] МОЛЧАНОВ И. Н., ЯКОВЛЕВ М. Ф., О двухшаговых итерационных методах решения одного класса несовместных систем линейных алгебраических уравнений, *Докл. АН СССР*, **214**, 6 (1974).

[14] НЕМЧИНОВ С. В., ЛИБОВ С. Л., Прямой метод повышенной точности решения краевых задач для уравнения Гельмгольца на сетке точек в прямоугольнике, *Ж. вычисл. матем. и матем. физ.*, 4, 4 (1964).
[15] ФАДДЕЕВ Д. К., О некоторых последовательностях полиномов, полезных для построения итерационных методов решения систем линейных алгебраических уравнений, *Вестник Ленингр. ун-та. Сер. матем.*, 7, 2 (1958).
[16] ФАДДЕЕВ Д. К., ФАДДЕЕВА В. Н., Вычислительные методы линейной алгебры. Записки научных семинаров ЛОМИ, т. 54, « Наука », 1975.
[17] ФАДДЕЕВ Д. К., ФАДДЕЕВА В. Н., КУБЛАНОВСКАЯ В. Н., Линейные алгебраические системы с прямоугольными матрицами. In : Вычислительные методы линейной алгебры, М., « Наука », 1968.
[18] BAUER F. L., FIKE C. T., Norms and exclusion theorems, *Numer. Math.*, 2, 3 (1960).
[19] BELLMAN R., *Introduction to matrix analysis*, New York, 1960.
[20] D'YAKONOV E. G., On the solution of some elliptic difference equations, *J. Inst. Math. and Appl.*, 7 (1971).
[21] EBERLEIN P., A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix, *J. Soc. Industr. Appl. Math.*, 10 (1962).
[22] FADDEEV D. K., FADDEEVA V. N., *Computational methods of linear algebra*, Freeman, San-Francisco, 1963.
[23] FLANDERS D. A., SHORTLEY G., Numerical determination of fundamental modes, *J. Appl. Phys.*, 21 (1950).
[24] FORSYTHE G., MOLER C., *Computer solution of linear algebraic systems*, Englewood Cliffs, New York, 1967.
[25] FRENCIS J., The QR-transformation, parts I, II, *Computer J.*, 4 (1961, 1962).
[26] GANTMAHER F. R., *La théorie des matrices*, *Paris*, Dunod, 1966.
[27] KELLOGG R., NODERER L., Sealed iterations and linear equations, *SIAM J.*, 8, 4 (1960).
[28] KUZNETSOV Y. A., Iterative Methods for Solution of Noncompatible Systems of Linear Equations. *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, 134 (1976).
[29] KUZNETSOV Yu.A., MARCHUK G. I., Stationary iterative methods for the solution of systems of linear equations with singular matrices, Gatlinburg VI, Symposium on Numerical Algebra. (Hoffen am See, 1974), Münich, 1974.
[30] MAREK I., On iteration of linear bounded operators and the convergence of Kellog's iteration process, *Cech. Mat. J.*, 12 (1962).
[31] RUTISHAUSER H., The Jacobi method for real symmetric matrices, *Numer. Math.*, 9, 1 (1966).
[32] STIEFEL E., Kernel polynomials in linear algebra and their numerical applications. *N.B.S., Appl. Math., Ser.* 49, 1 (1958).
[33] STRANG G., *Linear Algebra and its Applications*, N.Y., Academic Press, 1976.
[34] TRAUB J., *Iterative methods for the solution of equations*, Englewood Cliffs, Prentice Hall, 1964.
35] WILKINSON J., *The algebraic eigenvalue problem*, Oxford, Clarendon Press, 1965.
[36] WILKINSON J., REINSCH C., *Linear Algebra*, Springer-Verlag, Berlin, Heidelberg, 1971.
[37] YOUNG D. M., *Iterative Solution of Large Linear Systems*, Acad. Press, London, 1971.

IX. Méthodes spectrales d'optimisation des processus itératifs

[1] АБРАМОВ А. А., Об одном способе ускорения итерационных процессов, *Докл. АН СССР*, 74, 6 (1950).

[2] БАХВАЛОВ Н. С., О сходимости одного релаксационного метода при естественных ограничениях на эллиптический оператор, *Ж. вычисл. матем. и матем. физ.*, 6, 5 (1966).

[3] ВОРОБЬЕВ Ю. В., Случайный итерационный процесс, *Ж. вычисл. матем. и матем. физ.*, 4, 6 (1964); 5, 5 (1965).

[4] ГАВУРИН М. К., а) Применение полиномов наилучшего приближения для улучшения сходимости итерационных процессов, *Усп. матем. наук*, V, 3 (1950).
б) Нелинейные функциональные уравнения и непрерывные аналоги итерационных методов, *Изв. высш. учебн. заведений. Математика*, 5, 6 (1958).

[5] ДЬЯКОНОВ Е. Г., О построении итерационных методов на основе использования операторов эквивалентных по спектру, *Ж. вычисл. матем. и матем. физ.*, 6, 1, 4 (1966).

[6] ЗОЛОТАРЕВ Е. И., Приложение эллиптических функций к вопросам о функциях, наименее и наиболее отклоняющихся от нуля, *Записки Российской Академии наук*, 1877.

[7] ИВАНОВ В. К., О сходимости итерационных процессов при решении систем линейных алгебраических уравнений, *Изв. АН СССР, Сер. матем.*, 4 (1939).

[8] ЛЕБЕДЕВ В. И., а) Об итерационных методах решения операторных уравнений со спектром, лежащим на нескольких отрезках, *Ж. вычисл. матем. и матем. физ.*, 9, 4 (1969).
б) О построении операции P в KP-методе, *Ж. вычисл. матем. и матем. физ.*, 9, 4 (1969).

[9] ЛЕБЕДЕВ В. И., ФИНОГЕНОВ С. А., а) О порядке выбора итерационных параметров в чебышевском циклическом итерационном методе, *Ж. вычисл. матем. и матем. физ.*, 11, 2 (1971).
б) Решение проблемы упорядочения параметров в чебышевских итерационных методах, *Ж. вычисл. матем. и матем. физ.*, 13, 1 (1973).

[10] МАРЧУК Г. И., САРБАСОВ К. Е., Об одном методе решения стационарной задачи, *Докл. АН СССР*, 182, 1 (1968).

[11] ФЕДОРЕНКО Р. П., а) Релаксационный метод решения разностных эллиптических уравнений, *Ж. вычисл. матем. и матем. физ.*, 1, 5 (1961).
б) О скорости сходимости одного итерационного процесса, *Ж. вычисл. матем. и матем. физ.*, 4, 3 (1964).

[12] COLLATZ L., Fehlerabschätzung für das Iterationsverfahren zur Auflösung linear Gleichungssysteme, *Z. Angew. Math. Mech.*, 22 (1942).

[13] GOLUB G. H., VARGA R. S., Chebyshev semi-iterative methods, successive over-relaxation iterative methods and second order Richardson iterative methods, parts I, II, *Numer. Math.*, 3, 2 (1961).

[14] JUNCOSA M. L., MILLIKEN T. M., On the increase of convergence rates of relaxation procedures for elliptic partial difference equations, *J. Assoc. Comput. Mach.*, 7, 1 (1960).

[15] LANCZOS C., a) An iteration method for the solution of the eigenvalue problem of linear integral and differential operators, *J. Res. Nat. Bur. Stand.*, 45, 1 (1950).
b) Chebyshev polynomials in the solution of large-scale linear systems, *Proc. Assoc. Comput. Math.*, Toronto Meeting (September 1952).

[16] OSTROWSKI A. M., On the linear iteration procedures for symmetric matrices, Univ. Roma, Inst. Naz., *Alta Mat. Rend. Mat. e Appl.*, 14, 1-2 (1954).

[17] PETRYSHYN W., On a general iterative method for the approximate solution of linear operator equations, *Math. Comp.*, 17, 1 (1963).

[18] REICH E., On the convergence of the classical iterative method of solving linear simultaneous equations, *Ann. Math. Statist.*, **20**, 3 (1949).

X. Méthode de surrelaxation

[1] ФАДДЕЕВ Д. К., К вопросу о верхней релаксации при решении систем линейных уравнений, *Изв. высш. учебн. заведений. Математика*, **5** (1958).

[2] BROYDEN C. G., a) Some generalizations of the theory of successive over-relaxation, *Numer. Math.*, 6, 4 (1964).
b) On convergence criteria for the method of successive over-relaxation, *Math. Comput.*, **18**, 85 (1964).

[3] EVANS D. J., Note on the line over-relaxation factor for small mesh size, *Comput. J.*, 5, 1 (1962).

[4] EVANS D. J., FORINGTON C. V., An iterative process for optimizing symmetric successive over-relaxation, *Comput. J.*, 6, 3 (1963).

[5] FRANKEL S. P., Convergence rates of iterative treatments of partial differential equations, *Math. Tables and Other Aids Comput.*, 4, 30 (1950).

[6] GARABEDIAN P., Estimation of the relaxation factor for small mesh size, *Math. Tables and Other Aids Comput.*, **10**, 56 (1956).

[7] GASTINEL N., Sur le meilleur choix des paramètres de sur-relaxation (procédé de Peaceman-Rachford), *Chiffres*, 5, 2 (1962).

[8] GOLUB G. H., The use of Chebyshev matrix polynomials in the iterative solution of linear equations compared with the method of successive over-relaxation. (Doct. Thesis), Univ. of Illinois, **133** (1959).

[9] HAGEMAN L. A., KELLOGG R. B., Estimating optimum over-relaxation parameters, *Math. Comput.*, **22**, 101 (1968).

[10] LINN M. S., On the round-off error in the method of successive over-relaxation, *Math. Comput.*, **18**, 85 (1964).

[11] OSTROWSKI A. M., On over- and under-relaxation in the theory of the cyclic single step iteration, *Math. Tables and Other Aids Comput.*, **7**, 43 (1953).

[12] PETRYSHYN W., a) The generalized over-relaxation method for the approximate solution of operator equations in Hilbert space, *SIAM J.*, **10**, 4 (1962).
b) On the extrapolated Jacobi or simultaneous displacements method in the solution of matrix and operator equations, *Math. Comput.*, **19**, 89 (1965).

[13] SHELDON J., On the numerical solution of elliptic difference equation, *Math. Tables and other Aids Comput.*, 9 (1955).

[14] TEE G. J., a) Eigenvectors of the successive over-relaxation process, and its combination with Chebyshev semi-iteration, *Comput. J.*, 6, 3 (1963).
b) Correction to « Eigenvectors of the successive over-relaxation process, and its combination with Chebyshev semi-iteration », *Comput. J.*, 7, 1 (1964).

[15] VARGA R. S., a) *P*-cyclic matrices : a generalization of Young-Frankel successive over-relaxation scheme, *Pacific J. Math.*, 9 (1959).
b) Orderings of the successive over-relaxation scheme, *Pacific J. Math.*, 9 (1959).

[16] YOUNG D. M., a) Iterative methods for solving partial difference equations of elliptic type, *Trans. Amer. Math. Soc.*, **76** (1954).
b) A bound for the optimum relaxation factor for the successive over-relaxation method, *Numer. Math.*, **16**, 5 (1971).
c) Convergence properties of the symmetric and unsymmetric successive over-relaxation methods and related methods, *Math. Comput.*, **24**, 112 (1971).

XI. Méthodes du type gradient

[1] БИРМАН М. Ш., Некоторые оценки для метода наискорейшего спуска, *Усп. матем. наук*, V, 3 (1950).

[2] ГОДУНОВ С. К., ПРОКОПОВ Г. П., а) Вариационный подход к решению больших систем линейных уравнений, возникающих в сильно эллиптических задачах. Prépring Ин-та прикл. матем. АН СССР, М., 1968.
б) О решении разностного уравнения Лапласа, *Ж. вычисл. матем. и матем. физ.*, 9, 2 (1969).

[3] ГОРБЕНКО Н. И., ИЛЬИН В. П., О градиентных методах переменных направлений. In : Некоторые проблемы вычислительной и прикладной математики, Новосибирск, « Наука », 1975.

[4] КАНТОРОВИЧ Л. В., О методе наискорейшего спуска, *Докл. АН СССР*, 56, 3 (1947).

[5] КРАСНОСЕЛЬСКИЙ М. А., КРЕЙН С. Г., Итеративный процесс с минимальными невязками, *Матем. сб.*, **31** (1952).

[6] КУЗНЕЦОВ Ю. А., а) К теории итерационных процессов, *Докл. АН СССР*, **184**, 2 (1969).
б) О симметризации итерационных процессов. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1969.
в) Некоторые вопросы теории и приложений итерационных методов. (Thèse cand.), Новосибирск, 1969.

[7] МАРЧУК Г. И., КУЗНЕЦОВ Ю. А., а) К вопросу об оптимальных итерационных процессах, *Докл. АН СССР*, **181**, 6 (1968).
б) Некоторые вопросы теории многошаговых итерационных процессов. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1969.
в) К решению систем линейных уравнений итерационными методами. In : Вопросы точности и эффективности вычислительных алгоритмов, вып. 1, Киев, Изд. Ин-та кибернетики АН УССР, 1969.

[8] САМОКИШ Б. А., Исследование быстроты сходимости метода наискорейшего спуска, *Усп. матем. наук*, **XII**, 1 (1957).

[9] ФРИДМАН В. М., Новые методы решения линейного операторного уравнения, *Докл. АН СССР*, **128**, 3 (1959).

[10] DANIEL J. W., a) The conjugate gradient method for linear and nonlinear operator equations, *SIAM J. Numer. Anal.*, **4**, 1 (1967).
b) Convergence of the conjugate gradient method with computationally convenient modifications, *Numer. Math.*, **10**, 2 (1967).

[11] FORSYTHE G. E., On the asymptotic directions of the s-dimensional optimum gradient method, *Numer. Math.*, **11**, 1 (1968).

[12] FORSYTHE A. I., FORSYTHE G. E., Punched card experiments with accelerated gradient methods for linear equations. Contributions to the solution of linear equations and the determination of eigenvalues, *N.B.S. Appl. Math.*, Ser. 39, (1954).

[13] FORSYTHE G. E., MOTZKIN T. S., a) Asymptotic properties of the optimum gradient method, *Bull. Amer. Math. Soc.*, **57**, 2 (1951).
b) Acceleration of the optimum gradient method, *Bull. Amer. Math. Soc.*, **57**, 4 (1951).

[14] HESTENES M. R., STIEFEL E., Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Stand.*, **49** (1952).

[15] LANCZOS C., Solution of the system of linear equations by minimized iterations, *J. Res. Nat. Bur. Stand.*, **49**, 1 (1952).

XII. Factorisation (balayage)

[1] АБРАМОВ А. А., АНДРЕЕВ В. Б., О применении метода прогонки к нахождению периодических решений дифференциальных и разностных уравнений, *Ж. вычисл. матем. и матем. физ.*, **3**, 2 (1963).

[2] АПОСТОЛОВ Т., ВОЗНИЦКИ З., Диффузионная двумерная программа HEXAGA II для многогрупповых расчетов гексогональных решеток, *Атомн. энергия*, 38, 6 (1975).
[3] БУЛЕЕВ Н. И., а) Численный метод решения двумерных и трехмерных уравнений диффузии, *Матем. сб.*, 51, 2 (1960).
б) Метод неполной факторизации для решения двумерных и трехмерных уравнений типа диффузии, *Ж. вычисл. матем. и матем. физики*, 10, 4 (1970).
в) О численном решении двумерных уравнений эллиптического типа. In : Численные методы механики сплошной среды, Новосибирск, ВЦ СО АН СССР, 6, 3 (1975).
[4] ВЛАДИМИРОВ В. С., Приближенное решение одной краевой задачи для дифференциального уравнения второго порядка, *Прикл. матем. и мех.*, 19, 3 (1955).
[5] ГЕЛЬФАНД И. М., ЛОКУЦИЕВСКИЙ О. В., Метод прогонки для решения разностных уравнений. In : С. К. Годунов, В. С. Рябенький, *Введение в теорию разностных схем*, М., Физматгиз, 1962.
[6] ГИНКИН В. П., Новая схема неполной факторизации для решения двумерных уравнений эллиптического типа. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1974.
[7] ГОДУНОВ С. К., Метод ортогональной прогонки для решения систем разностных уравнений, *Ж. вычисл. матем. и матем. физ.*, 2, 6 (1962).
[8] ДЕГТЯРЕВ Л. М., ФАВОРСКИЙ А. П., а) Потоковый вариант метода прогонки, *Ж. вычисл. матем. и матем. физ.*, 8, 3 (1968).
б) Потоковый вариант метода прогонки для разностных задач с сильно меняющимися коэффициентами, *Ж. вычисл. матем. и матем. физ.*, 9, 1 (1969).
[9] ОГНЕВА В. В., Метод « прогонки » для решения разностных уравнений, *Ж. вычисл. матем. и матем. физ.*, 7, 4 (1967).
[10] РУСАНОВ В. В., Об устойчивости метода матричной прогонки. In : Вычислительная математика, № 6, М., 1960.
[11] СОФРОНОВ И. Д., а) О методе прогонки для решения краевых задач для разностных уравнений, *Ж. вычисл. матем. и матем. физ.*, 4, 2 (1964).
б) Разностная схема с диагональными направлениями прогонок для решения уравнения теплопроводности, *Ж. вычисл. матем. и матем. физ.*, 5, 2 (1965).
[12] ФАГЕ М. К., О методе прогонки, *Докл. АН СССР*, 191, 2 (1970).
[13] BUZBEE B., GOLUB G., NIELSON C., On direct methods for solving Poisson's equations, *SIAM J. Numer. Anal.*, 7, 4 (1970).
[14] INCE E. L., *Ordinary Differential Equations*, Longmans Green and Co., London, 1927.
[15] OLIPHANT T. A., An implicit numerical method for solving two-dimensional time dependent diffusion problems, *Quarterly of Applied Mathematics*, XIX, 3 (1961).

XIII. Transformation de Fourier rapide

[1] КУЗНЕЦОВ Ю. А., МАЦОКИН А. М., Решение уравнения Гельмгольца методом фиктивных областей. In : Вычислительные методы линейной алгебры, Новосибирск, ВЦ СО АН СССР, 1972.
[2] НЕМЧИНОВ С. В., О применении метода сеток к решению краевых задач для уравнений в частных производных с периодическими краевыми условиями. In : Динамическая метеорология, Ташкент, « Наука », 1965.

[3] BINGHAM C., GODFRYE M. D., TUKEY J., Modern techniques of power spectrum estimation, *IEEE Trans., Audio and Electroacoustics*, AU-15 (1967).
[4] COOLEY J. W., LEWIS P. A., WELCH P. D., The fast Fourier transform algorithm and its applications, IBM Research Paper RC-1743, Feb. (1967).
[5] COOLEY J. W., TUKEY J. W., An algorithm for the machine calculation of complex Fourier series, *Math. Comput.*, 19, 90 (1965).
[6] GOLD B., RADER C. M., *Digital processing of signal*, McGraw-Hill, N.Y., 1969.
[7] HELMS R. D., Fast Fourier method for computing difference equations and simulating filters, *IEEE Trans., Audio and Electroacoustics*, AU-15 (1967).
[8] HOCKNEY R. W., A fast direct solution of Poisson's equation using Fourier analysis, *J. Assoc. Comp. Mach.*, 12, 1 (1965).
[9] KLAUDER J. R., PRICE A. C., DARLINGTON S., ALBERSHEIM W. J., The theory and design of chirp radars, *The Bell System Tech., J.*, 39, 4 (1960).
[10] TUKEY J. W., An introduction to the calculations of numerical spectrum analysis. In : « Spectral Analysis in Time Series », Bern Harris, L.D. Wiley, N.Y., 1967.

XIV. Interpolation à l'aide de fonctions splines

[1] АНАНЬИН А. З., СМЕЛОВ В. В., ВАСИЛЕНКО В. А., Эффективный способ преобразования вариационной задачи сглаживания к линейной алгебраической системе. Préprint № 28, Новосибирск, ВЦ СО АН СССР, 1976.
[2] ВАСИЛЕНКО В. А., а) Сходимость сплайнов в гильбертовом пространстве. In : Численные методы механики сплошной среды, 3, 3 (1972).
б) Сходимость операторных интерполирующих сплайнов. In : Вариационно-разностные методы в математической физике, Новосибирск, ВЦ СО АН СССР, 1973.
в) Сглаживающие сплайны на подпространствах и теоремы компактности. In : Численные методы механики сплошной среды, Новосибирск, ВЦ СО АН СССР, 5, 5 (1974).
г) Конечномерная аппроксимация в методе наименьших квадратов. In : Вариационно-разностные методы математической физики, Новосибирск, ВЦ СО АН СССР, 2 (1975).
[3] ВАСИЛЕНКО В. А., ПЕРЕЛОМОВ Е. М., Сплайн-интерполяция в прямоугольной области с хаотически расположенными узлами. In : Машинная графика и ее применение, Новосибирск, ВЦ СО АН СССР, 1973.
[4] ГОРДОНОВА В. И., МОРОЗОВ В. А., Численные алгоритмы выбора параметра в методе регуляризации, *Ж. вычисл. матем. и матем. физ.*, 13, 3 (1973).
[5] ЗАВЬЯЛОВ Ю. С., а) Интерполирование кубическими многозвенниками. In : Вычислительные системы, вып. 38, Новосибирск, ИМ СО АН СССР, 1970.
б) Интерполирование бикубическими многозвенниками. In : Вычислительные системы, вып. 38, Новосибирск, ИМ СО АН СССР, 1970.
в) Экстремальное свойство кубических многозвенников и задача сглаживания. In : Вычислительные системы, вып. 42, Новосибирск, ИМ СО АН СССР, 1970.
г) Интерполирование мультикубическими сплайнами. In : Вычислительные системы, вып. 65, Новосибирск, ИМ СО АН СССР, 1975.

[6] ЛЕБЕДЕВ В. И., Об одном способе интерполяции в n-мерном пространстве по произвольным узлам и некоторых квадратурных формулах. Préprint ВЦ СО АН СССР, вып. 10, Новосибирск, 1975.
[7] МИХАЛЕВИЧ Ю. И., ОМЕЛЬЧЕНКО О. К., Процедуры кусочно-полиномиальной интерполяции функций одной и двух переменных. Новосибирск, ВЦ СО АН СССР, 1970.
[8] МОРОЗОВ В. А., а) О выборе параметра при решении функциональных уравнений методом регуляризации, *Докл. АН СССР*, 175, 6 (1967).
б) Теория сплайнов и задача устойчивого вычисления значений неограниченного оператора, *Ж. вычисл. матем. и матем. физ.*, 11, 3 (1971).
[9] ПИВОВАРОВА Н. Б., ПУХНАЧЕВА Т. П., Сглаживание экспериментальных данных локальными сплайнами. Préprint ВЦ СО АН СССР, вып. 9, Новосибирск, 1975.
[10] РЯБЕНЬКИЙ В. С., Локальные формулы гладкого восполнения и гладкой интерполяции функций по их значениям в узлах неравномерной прямоугольной сетки. Préprint № 21 ИПМ СССР, 1974.
[11] ЦЕЦОХО В. А., БЕЛОНОСОВ А. С., БЕЛОНОСОВА А. В., Об одном методе r-гладкого приближения функций многих переменных. Préprint ВЦ СО АН СССР, вып. 8, Новосибирск, 1974.
[12] ЯНЕНКО Н. Н., КВАСОВ Б. И., Итерационный метод построения поликубических сплайн-функций. In : Численные методы механики сплошной среды, Новосибирск, ВЦ СО АН СССР, 1, 3 (1970).
[13] ANSELONE P. M., LAURENT P. J., A general method for construction of interpolating or smoothing spline-functions, *Numer. Math.*, 12, 1 (1968).
[14] ATTEIA M., Généralisation de la définition et des propriétés des « spline fonctions », *C.R. Acad. Sci., Paris*, 260, 13 (1965).
[15] BIRKHOFF G., GARABEDIAN P., Smooth surface interpolation, *J. Math. Phys.*, 39, 3 (1960).
[16] BOOR C. de, Bicubic spline interpolation, *J. Math. Phys.*, 41, 2 (1962).
[17] HOLLADAY J. C., Smoothest curve approximation, *Math. Tables and other Aids Comput.*, 11, 60 (1957).
[18] JOLY J. L., Convergence des fonctions-spline, Vème Congrès de l'AFIRO, Lille, 1966.
[19] REINSCH C. H., Smoothing by spline functions, *Numer. Math.*, 10, 4 (1967).
[20] SCHOENBERG I. J., Contribution to the problem of approximation of equidistant data by analitic functions. Parts A and B, *Quart. Appl. Math.*, 4 (1946).
[21] SCHUMAKER L. L., *Approximation by splines : Theory and applications of spline functions*, Academic Press, New York-London, 1969.
[22] WALSH J. L., AHLBERG J. H., NILSON E. N., Best approximation properties of the spline fit, *J. Math. Mech.*, 11, 2 (1962).

XV. Méthodes de décomposition

[1] АНДРЕЕВ В. Б., О разностных схемах с расщепляющимся оператором для общих p-мерных параболических уравнений второго порядка со смешанными производными, *Ж. вычисл. матем. и матем. физ.*, 7, 2 (1967).
[2] БАГРИНОВСКИЙ К. А., ГОДУНОВ С. К., Разностные методы для многомерных задач, *Докл. АН СССР*, 115, 3 (1957).
[3] БУЛЕЕВ Н. И., Численный метод решения двумерных и трехмерных уравнений диффузии, *Матем. сб.*, 51, 2 (1960).
[4] ВОРОБЬЕВ Ю. В., Случайный итерационный процесс в методе переменных направлений, *Ж. вычисл. матем. и матем. физ.*, 8, 3 (1968).
[5] ДЬЯКОНОВ Е. Г., а) Метод переменных направлений решения систем

конечно-разностных уравнений, *Докл. АН СССР*, **138**, 2 (1961).
б) О некоторых разностных схемах для решения краевых задач, *Ж. вычисл. матем. и матем. физ.*, **2**, 1 (1962).
в) Разностные схемы с расщепляющимся оператором для многомерных стационарных задач, *Ж. вычисл. матем. и матем. физ.*, **2**, 4 (1962).
г) Решение некоторых многомерных задач математической физики при помощи сеток. (Thèse cand.), М., 1962.

[6] ДЬЯКОНОВ Е. Г., ЛЕБЕДЕВ В. И., Метод расщепления для третьей краевой задачи. In : Вычислительные методы и программирование, вып. IV, М., Изд. Моск. ун-та, 1967.

[7] ИЛЬИН В. П., а) О расщеплении разностных уравнений параболического и эллиптического типов, *Сиб. матем. ж.*, **VI**, 1 (1965).
б) О явных схемах переменных направлений, *Изв. Сиб. отд. АН СССР. Сер. техн. н.*, **13**, 3 (1967).

[8] КОНОВАЛОВ А. Н., а) Метод дробных шагов решения задачи Коши для многомерного уравнения колебаний, *Докл. АН СССР*, **147**, 1 (1962).
б) Применение метода расщепления к численному решению динамических задач теории упругости, *Ж. вычисл. матем. и матем. физ.*, **4**, 4 (1964).
в) *Задачи фильтрации многофазной несжимаемой жидкости*, Новосибирск, НГУ, 1972.

[9] МАРЧУК Г. И., СУЛТАНГАЗИН У. М., К обоснованию метода расщепления для уравнения переноса излучения, *Ж. вычисл. матем. и матем. физ.*, 5, 5 (1965).

[10] МАРЧУК Г. И., ЯНЕНКО Н. П., Применение метода расщепления (дробных шагов) для решения задач математической физики. In : Некоторые вопросы вычислительной и прикладной математики, Новосибирск, « Наука », 1966.

[11] САМАРСКИЙ А. А., а) Об одном экономичном разностном методе решения многомерного параболического уравнения в произвольной области, *Ж. вычисл. матем. и матем. физ.*, **2**, 5 (1962).
б) О сходимости метода дробных шагов для уравнения теплопроводности, *Ж. вычисл. матем. и матем. физ.*, **2**, 6 (1962).
в) Локально-одномерные разностные схемы на неравномерных сетках, *Ж. вычисл. матем. и матем. физ.*, **3**, 3 (1963).
г) Об одном экономичном алгоритме численного решения систем дифференциальных и алгебраических уравнений, *Ж. вычисл. матем. и матем. физ.*, **4**, 3 (1964).
д) Экономичные разностные схемы для гиперболической системы уравнений со смешанными производными и их применение для уравнений теории упругости, *Ж. вычисл. матем. и матем. физ.*, 5, 1 (1965).
е) Аддитивные схемы. In : Тезисы докладов на Международном съезде математиков в Москве, 1966.

[12] ФРЯЗИНОВ И. В., О разностных схемах для уравнения Пуассона в полярной, цилиндрической и сферической системах координат, *Ж. вычисл. матем. и матем. физ.*, **11**, 5 (1971).

[13] ЯНЕНКО Н. Н., а) Об одном разностном методе счета многомерного уравнения теплопроводности, *Докл. АН СССР*, **125**, 6 (1959).
б) Об экономичных неявных схемах (метод дробных шагов), *Докл. АН СССР*, **134**, 5 (1960).
в) О неявных разностных методах счета многомерного уравнения теплопроводности, *Изв. высш. учебн. заведений. Математика*, 4 (23) (1961).
г) О сходимости метода расщепления для уравнения теплопроводности с переменными коэффициентами, *Ж. вычисл. матем. и матем. физ.*, **2**, 5 (1962).
д) О слабой аппроксимации систем дифференциальных уравнений, *Сиб. матем. ж.*, **V**, 6 (1964).

[14] ЯНЕНКО Н. Н., ДЕМИДОВ Г. В., Метод слабой аппроксимации как конструктивный метод построения решения задачи Коши. Iп : Некоторые вопросы вычислительной и прикладной математики, Новосибирск, « Наука », 1966.
[15] BAKER G. A., An implicit numerical method for solving the n-dimensional heat equation, *Quart. Appl. Math.*, **17**, 4 (1960).
[16] BAKER G. A., OLIPHANT T. A., An implicit numerical method for solving the two-dimensional heat equation, *Quart. Appl. Math.*, **17**, 4 (1960).
[17] BENSOUSSAN A., Pure decentralization for interrelated payoffs. Symposium on Optimization, Los Angeles, 1971.
[18] BENSOUSSAN A., LIONS J. L., TEMAM R., Sur les méthodes de décomposition, de décentralisation et de coordination et applications. Sur les méthodes numériques en sciences physiques et économiques, Dunod, Paris, 1974.
[19] BIRKHOFF G., VARGA R. S., Implicit alternating direction method, *Trans. Amer. Math. Soc.*, **92**, 1 (1959).
[20] BIRKHOFF G., VARGA R. S., YOUNG D., Alternating direction implicit methods, *Advanced in Comp.*, **3**, Academic Press, New York-London, 1962.
[21] DOUGLAS J., GUNN J. E., a) Two high-order correct difference analogues for the equation of multi-dimensional heat flow, *Math. Comput.*, **17**, 81 (1963).
b) A general formulation of alternating direction methods. Part I. Parabolic and hyperbolic problems, *Numer. Math.*, **6**, 5 (1964).
[22] DOUGLAS J., JONES B. F., On predictor-corrector methods for nonlinear parabolic differential equations, *J. Soc. Industr. Appl. Math.*, **11**, 1 (1963).
[23] DOUGLAS J., KELLOGG R. B., VARGA R. S., Alternating direction methods for n-space variables, *Math. Comput.*, **17**, 83 (1963).
[24] DOUGLAS J., PEARCY C. M., On convergence of alternating direction procedures in the presence of singular operators, *Numer. Math.*, **5**, 2 (1963).
[25] DOUGLAS J., RACHFORD H., On the numerical solution of heat conduction problems in two- and three- space variables, *Trans. Amer. Math. Soc.*, **82**, 2 (1956).
[26] DUPONT T., A factorization procedure for the solution of elliptic difference equations, *SIAM J. Numer. Anal.*, **5**, 4 (1968).
[27] FAIRWEATHER G., MITCHELL A. R., Some computational results of an improved A.D.I. method for the Dirichlet problem, *Comput. J.*, **9**, 3 (1966).
[28] GUNN J. E., The solution of elliptic difference equations by semi-explicit iterative techniques, *SIAM J. Numer. Anal.*, **2**, 1 (1965).
[29] HUBBARD B. E., Alternating direction schemes for the heat equation in a general domain, *SIAM J. Numer. Anal.*, **2**, 3 (1966).
[30] KELLOGG R. B., a) Another alternating-direction implicit method, *J. Soc. Industr. Appl. Math.*, **11**, 4 (1963).
b) An alternating direction method for operator equations, *J. Soc. Industr. Appl. Math.*, **12**, 4 (1964).
[31] KELLOGG R. B., SPANIER J., On optimal alternating direction parameters for singular matrices, *Math. Comput.*, **19**, 91 (1965).
[32] KUZNETSOV B. G., Numerical methods for solving some problems of viscous liquid, *Fluid Dynamics Transactions*, v. 4, 1969.
[33] LEES M., a) Alternating direction methods for hyperbolic differential equations, *J. Soc. Industr. Appl. Math.*, **10**, 4 (1960).
b) Alternating direction and semi-explicit difference methods for parabolic partial differential equations, *Numer. Math.*, **3**, 5 (1961).
[34] MARCHUK G. I., On the theory of the splitting-up method. Numerical

solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.
[35] PEACEMAN D. W., RACHFORD H. H., The numerical solution of parabolic and elliptic differential equations, *SIAM J.*, 3, 1 (1955).
[36] TEMAM R., a) Sur la stabilité et la convergence de la méthode des pas fractionnaires, *Ann. mat. pura ed appl.*, IV, 79 (1968).
b) Quelques méthodes de décomposition en Analyse numérique. Actes du Congrès International des mathématiciens. 1/10 Septembre 1970 Nice, 3, Paris, 1971.
[37] VARGA R. S., Some results in approximation theory with applications to numerical analysis. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London 1971.
[38] WACHSPRESS E. L., a) Optimum alternating-direction-implicit iteration parameters for a model problem, *SIAM J.*, 10, 2 (1962).
b) *Iterative solution of elliptic systems and applications of the neutron diffusion equations of reactor physics*, Prentice-Hall, Englewood Cliffs, N.Y., 1966.
c) Extended application of alternating-direction-implicit iteration model problem theory, *SIAM J.*, 11, 3 (1963).
d) *SIAM J.*, 11, 4 (1963).
[39] WACHSPRESS E. L., HABETLER G. J., An alternating-direction-implicit iteration technique, *SIAM J.*, 8, 2 (1960).
[40] WIDLUND O., a) On the rate of an alternating direction-implicit method in a non-commutative case, *Math. Comput.*, 20, 96 (1966).
b) On the effects of scaling of the Peaceman-Rachford method, *Math. Comput.*, 25, 113 (1971).

XVI. Problèmes conditionnellement bien posés et certains problèmes inverses de la physique mathématique

[1] ИВАНОВ В. К., О некорректно поставленных задачах, *Матем. сб.*, 61, 2 (1963).
[2] КАДОМЦЕВ Б. Б., О функции влияния в теории переноса лучистой энергии, *Докл. АН СССР*, 113, 3 (1957).
[3] КРЕЙН С. Г., О классах корректности для некоторых задач, *Докл. АН СССР*, 114, 6 (1957).
[4] КРЕЙН С. Г., ПРОЗОРОВСКАЯ О. И., О приближенных методах решения некорректных задач, *Ж. вычисл. матем. и матем. физ.*, 3, 1 (1963).
[5] ЛАВРЕНТЬЕВ М. М., а) О задаче Коши для уравнения Лапласа, *Докл. АН СССР*, 102, 2 (1956).
б) О постановке некоторых некорректных задач математической физики. In : Некоторые вопросы вычислительной и прикладной математики, Новосибирск, « Наука », 1966.
[6] ЛАВРЕНТЬЕВ М. М., ВАСИЛЬЕВ В. Г., О постановке некоторых некорректных задач математической физики, *Сиб. матем. ж.*, VII, 3 (1966).
[7] ЛАНДИС Е. М., О некоторых свойствах решений эллиптических уравнений, *Докл. АН СССР*, 107, 4 (1956).
[8] ЛЕВИТАН Б. М., *Разложение по собственным функциям дифференциальных уравнений второго порядка*, М.-Л., Гостехиздат, 1950.
[9] ЛЕВИТАН Б. М., САРГСЯН И. С., *Введение в спектральную теорию*, М., « Наука », 1970.
[10] МАРЧУК Г. И., Уравнения для ценности информации с метеорологических спутников и постановка обратных задач, *Космические исследования*, 11, 3 (1964).

[11] МАРЧУК Г. И., АТАНБАЕВ С. А., Некоторые вопросы глобальной регуляризации, *Докл. АН СССР*, **190**, 3 (1970).
[12] МАРЧУК Г. И., ВАСИЛЬЕВ В. Г., О приближенном решении операторных уравнений первого рода, *Докл. АН СССР*, **199**, 4 (1970).
[13] МАРЧУК Г. И., ДРОБЫШЕВ Ю. П., Некоторые вопросы линейной теории измерений, *Автометрия*, **3** (1967).
[14] МАРЧУК Г. И., ОРЛОВ В. В., К теории сопряженных функций. In : Нейтронная физика, М., Госатомиздат, 1961.
[15] МЕРГЕЛЯН С. Н., Гармоническая аппроксимация и приближенное решение задачи Коши для уравнения Лапласа, *Усп. матем. наук*, **XI**, 5 (1956).
[16] МОРОЗОВ В. А., Методы решения неустойчивых задач, МГУ, 1967 (rotaprint).
[17] ТИХОНОВ А. Н., а) Об устойчивости обратных задач, *Докл. АН СССР*, **39**, 5 (1943).
б) О решении некорректно поставленных задач и методе регуляризации, *Докл. АН СССР*, **151**, 3 (1963).
в) О регуляризации некорректно поставленных задач, *Докл. АН СССР*, **153**, 1 (1963).
[18] УСАЧЕВ Л. Н., Уравнение для ценности нейтронов кинетического реактора и теория возмущения. In : Реакторостроение и теория реакторов, М., АН СССР, 1955.
[19] ФАДДЕЕВА В. Н., Сдвиг для систем с плохо обусловленными матрицами, *Ж. вычисл. матем. и матем. физ.*, 5, 5 (1965).
[20] ФРАНК Л. С., ЧУДОВ Л. А., Разностные методы решения некорректной задачи Коши. In : Численные методы в газовой динамике, М., МГУ, 1965.
[21] DOUGLAS J., On the relation between stability and convergence in the numerical solution of linear parabolic and hyperbolic differential equations, *J. Soc. Industr. Appl. Math.*, **4**, 1 (1956).
[22] FUKS K., Perturbation theory in neutron multiplication problems, *Proc. Phys. Soc.*, **62**, 791 (1949).
[23] JOHN F., Differential equations with approximate and improper data, *Lectures, N.Y. Univ.*, 1955.
[24] LATTÈS R., LIONS J. L., *Méthode de quasi-réversibilité et applications*, Paris, Dunod, 1967.
[25] LAVRENTIEV M. M., Numerical solution of conditionally properly posed problems. Numerical solution of partial differential equations — II. SYNSPADE 1970, Academic Press, New York-London, 1971.
[26] LAVRENTIEV M. M., ROMANOV V. G., VASILIEV V. G., Multidimensional inverse problems for differential equations, *Lectures Notes in Mathematics*, **167**, 1970, Springer-Verlag.
[27] MARCHUK G. I., Formulation of the theory of perturbations for complicated models, *Appl. Math. Optim.*, **2**, 1 (1975).
[28] TIKHONOV A., ARSÉNINE V., *Méthode de résolution de problèmes mal posés*, Ed. de Moscou, 1976.

XVII. Méthodes numériques en théorie du transport

[1] АГОШКОВ В. И., а) Вариационные методы в теории переноса. (Thèse cand.). Новосибирск, ВЦ СО АН СССР, 1975.
б) Выбор базисных функций при решении некоторых эллиптических уравнений. In : Вычислительная математика и программирование, Новосибирск, ВЦ СО АН СССР, 1974.
[2] БОГОЛЮБОВ Н. Н., *Проблемы динамической теории в статистической физике*, М., Гостехтеориздат, 1946.

[3] ВЛАДИМИРОВ В. С., а) Численное решение кинетического уравнения для сферы. In : Вычислительная математика, № 3, М., АН СССР, 1958.
б) Математические задачи односкоростной теории переноса частиц, *Труды Матем. ин-та АН СССР*, **61** (1961).
в) О некоторых вариационных методах приближенного решения уравнения переноса. In : Вычислительная математика, № 7, М., АН СССР, 1961.

[4] ГЕРМОГЕНОВА Т. А., а) О сходимости некоторых приближенных методов решения уравнения переноса, *Докл. АН СССР*, **181**, 3 (1968).
б) Обобщенные решения краевых задач для уравнения переноса, *Ж. вычисл. матем. и матем. физ.*, **9**, 3 (1969).

[5] ГОДУНОВ С. К., Использование интеграла энергии для оценки точности приближенных собственных значений, *Ж. вычисл. матем. и матем. физ.*, **11**, 5 (1971).

[6] ГОДУНОВ С. К., СУЛТАНГАЗИН У. М., О диссипативности граничных условий В. С. Владимирова для симметрической системы метода сферических гармоник, *Ж. вычисл. матем. и матем. физ.*, **11**, 3 (1971).

[7] ГОЛЬДИН В. Я., Квазидиффузионный метод решения кинетического уравнения, *Ж. вычисл. матем. и матем. физ.*, **4**, 6 (1964).

[8] КУЗНЕЦОВ Е. С., МАРЧУК Г. И., Вычислительные методы в теории переноса излучения. In : Труды IV Всесоюзного математического съезда, т. II, М., АН СССР, 1964.

[9] ЛЕБЕДЕВ В. И., а) О нахождении решений кинетических задач теории переноса. (Thèse doct.). Новосибирск, 1967.
б) О *KP*-методе и разностных схемах для кинетического уравнения. In : Вычислительные методы в теории переноса, М., Атомиздат, 1969.

[10] МАРЧУК Г. И., Численные методы расчета ядерных реакторов, М., Атомиздат, 1958.

[11] МАРЧУК Г. И., КОЧЕРГИН В. П., Эффективный метод решения двумерного уравнения диффузии для ячеек квадратной и шестиугольной формы, *Атомная энергия*, **18**, 6 (1965).

[12] МАРЧУК Г. И., ЛЕБЕДЕВ В. И., *Численные методы в теории переноса нейтронов*, М., Атомиздат, 1971.

[13] МАРЧУК Г. И., СУЛТАНГАЗИН У. М., а) О сходимости метода расщепления для уравнений переноса излучений, *Докл. АН СССР*, **161**, 1 (1965).
б) К вопросу о решении кинетического уравнения переноса методом расщепления, *Докл. АН СССР*, **163**, 4 (1965).

[14] МАРЧУК Г. И., ЯНЕНКО Н. Н., Решение многомерного кинетического уравнения методом расщепления, *Докл. АН СССР*, **157**, 6 (1964).

[15] НИКОЛАЙШВИЛИ Ш. С., а) Приближенное решение уравнения переноса методом моментов, *Атомная энергия*, **9**, 2 (1961).
б) О решении односкоростного уравнения переноса с использованием приближения Ивона-Мартенса, *Атомная энергия*, **20**, 4 (1966).

[16] СМЕЛОВ В. В., *Лекции по теории переноса нейтронов*, М., Атомиздат, 1972.

[17] СУЛТАНГАЗИН У. М., а) Дифференциальные свойства решений смешанной задачи Коши для нестационарного кинетического уравнения. Préprint ВЦ СО АН СССР, Новосибирск, 1971.
б) К обоснованию метода слабой аппроксимации для уравнения сферических гармоник. Préprint ВЦ СО АН СССР, Новосибирск, 1971.
в) Слабая сходимость метода сферических гармоник. Préprint ВЦ СО АН СССР, Новосибирск, 1971.

[18] ШИХОВ С. Б., Некоторые вопросы математической теории критического состояния реактора, *Ж. вычисл. матем. и матем. физ.*, **7**, 1 (1967).

[19] BARDOS P. G., Problèmes aux limites pour les équations aux dérivées

partielles du premier ordre à coefficients réels, *Ann. Sci. Ec. Norm. Sup.*, 4e série, **3**, 2 (1970).
[20] JORGENS K., An asymptotic expansion in the theory of neutron transport, *Comm. Pure Appl. Math.*, **11**, 2 (1958).
[21] MAREK I., On a problem of mathematical physics, *Apl. Mat.*, **11**, 89 (1966).

XVIII. Méthode de Monte-Carlo

[1] БУСЛЕНКО Н. П., ГОЛЕНКО Д. И., СОБОЛЬ И. М., СРАГОВИЧ В. Г., ШРЕЙДЕР Ю. А., *Метод статистических испытаний (метод Монте-Карло)*, М., Физматгиз, 1962.
[2] ВЛАДИМИРОВ В. С., СОБОЛЬ И. М., Расчет наименьшего характеристического числа уравнения Пайерлса методом Монте-Карло. In : Вычислительная математика, № 3, М., АН СССР, 1958.
[3] ГЕЛЬФАНД И. М., ФРОЛОВ А. С., ЧЕНЦОВ Н. Н., Вычисление континуальных интегралов методом Монте-Карло, *Изв. высш. учебн. заведений. Математика*, **5** (1958).
[4] ЕРМАКОВ С. М., *Метод Монте-Карло и смежные вопросы*, М., « Наука », 1971.
[5] ЕРМАКОВ С. М., ЗОЛОТУХИН В. Г., Полиномиальные приближения и метод Монте-Карло, *Теория вероятностей и ее применения*, 5, 4 (1960).
[6] ЕРМАКОВ С. М., МИХАЙЛОВ Г. А., *Курс статистического моделирования*, М., « Наука », 1976.
[7] КЕРТИС Д., Методы Монте-Карло для итерации линейных операторов, *Усп. матем. наук*, **XII**, 5 (1957).
[8] МАРЧУК Г. И., МИХАЙЛОВ Г. А., НАЗАРАЛИЕВ М. А., ДАРБИНЯН Р. А., КАРГИН Б. А., ЕЛЕПОВ Б. С., *Метод Монте-Карло в атмосферной оптике*, Новосибирск, « Наука », 1976.
[9] МИХАЙЛОВ Г. А., *Некоторые вопросы теории методов Монте-Карло*, Новосибирск, « Наука », 1974.
[10] СОБОЛЬ И. М., *Численные методы Монте-Карло*, М., « Наука », 1973.
[11] FANO U., SPENCER L. V., BERGER M. B., *Penetration and Diffusion of X Rays*. Handbuch du Physik. Band. XXXVIII/2. Neutronen und Verwandte Gammastralprobleme, Berlin-Göttingen-Heidelberg, 1959.
[12] METROPOLIS N., ULAM S., The Monte-Carlo method, *J. Amer. Stat. Assoc.*, **44**, 247 (1949).
[13] SPANIER J., GELBARD E., *Monte-Carlo principles and Neutron Transport Problems*, Addison-Wesley Publ. Co., 1969.

XIX. Méthode de grosses particules

[1] БЕЛОЦЕРКОВСКИЙ О. М., ДАВЫДОВ Ю. М., Нестационарный метод « крупных частиц » для газодинамических расчетов, *Ж. вычисл. матем. и матем. физ.*, **11**, 1 (1971).
[2] ВЕДЕШКИНА К. А., ЛЕВИНА З. Ф., ЛОМНЕВ С. П., ПРУДКОВСКИЙ Г. П., РАСТОПЧИНА Т. В., РУБЕН Г. В., ЮРЧЕНКО В. В., *Решение задач методом « крупных частиц »*, М., ВЦ АН СССР, 1970.
[3] ДЬЯЧЕНКО В. Ф., Об одном новом методе численного решения нестационарных задач газовой динамики с двумя пространственными переменными, *Ж. вычисл. матем. и матем. физ.*, 5, 4 (1965).
[4] ЯНЕНКО Н. Н., АНУЧИНА Н. Н., ПЕТРЕНКО В. Е., ШОКИН Ю. И., О методах расчета задач газовой динамики с большими деформациями. In : Численные методы механики сплошной среды, **1**, 1 (1970).
[5] HARLOW F. H., The particle in cell method for numerical solution of problems in fluid dynamics, Proc. Symp. in Applied Math. A.M.S., **XV** (1963).

XX. Méthodes d'optimisation des algorithmes

[1] БАБУШКА И., СОБОЛЕВ С. Л., Оптимизация численных методов, *Apl. Mat.*, **10**, 2 (1965).
[2] БАХВАЛОВ Н. С., Об оптимальных методах решения задач, *Apl. Mat.*, **13**, 1 (1968).
[3] ВИНОГРАДОВ И. М., К вопросу об оценке тригонометрических сумм, *Изв. АН СССР. Сер. матем.*, **29**, 3 (1965).
[4] КОЛМОГОРОВ А. Н., Дискретные автоматы и конечные алгоритмы. In : Труды IV Всесоюзного математического съезда, т. I, М., АН СССР, 1963.
[5] КОРОБОВ Н. М., Вычисление кратных интегралов методом оптимальных коэффициентов, *Вестн. Моск. ун-та. Сер. матем.*, **4**, 1959.
[6] МОИСЕЕВ Н. Н., Численные методы, использующие варьирование в пространстве состояний и некоторые вопросы управления большими системами. In : Тезисы докладов Международного конгресса математиков, М., 1966.
[7] МОИСЕЕВ Н .Н., КРАСОВСКИЙ Н. Н., Теория оптимальных управляемых систем, *Изв. АН СССР. Техн. кибернетика*, 5 (1967).
[8] ЧЕРНОУСЬКО Ф. Л., БАНИЧУК Н. В., ПЕТРОВ В. М., Численные решения вариационных и краевых задач методом локальных вариаций, *Ж. вычисл. матем. и матем. физ.*, **6**, 6 (1966).
[9] DAHLQUIST G., Convergence and stability in the numerical integration of ordinary differential equations, *Math. Scand.*, **4**, 1 (1956).
[10] MOORE R., *Interval analysis*, Prentice-Hall, 1966.
[11] NICKEL K., a) Über die Notwendigkeit einer Fehlerschranken-Arithmetic für Rechnenautomaten, *Numer. Math.*, **9**, 1 (1966).
b) Bericht über neue Kalsruher Ergebnisse bei der Fehlererfassung von numerischen Prozessen, *Apl. Mat.*, **13**, 2 (1968).

INDEX

Analyse des erreurs d'arrondi 394
Approximation
  d'un problème 31
  r-régulière 137

Cône 113
Condition
  de Courant 351
  d'ellipticité 105
  de stabilité numérique 355
Conditions aux limites
  essentielles 64
  naturelles 64
Critère
  des résidus normés 147
  spectral de stabilité 39

Domaine d'un opérateur 14

Ensembles de programmes standards 395
Equation
  de la chaleur 343
  de continuité 359
  de diffusion 273
  d'Euler 58
  d'évolution 34
  de Helmholtz 212
  hyperbolique 262
  du mouvement 353, 359, 363
  — sous forme conservative 361
  de Poisson 328
  de transport 369
  des vibrations 262, 348
Espace
  de Banach 115
  hilbertien 13
  $L_2(D)$ 13
  de Sobolev 16

Extrapolation
  à la limite 268
  de Richardson 268

Famille régularisante au sens de Tikhonov 295
Fonction
  de base 103
  discrète 28
  d'influence (d'importance) 312
Fonction spline 127
  bicubique 146
  cubique 127
  d'interpolation 144
  de lissage (d'ajustement) 154
  relative à un sous-espace 146
Fonctionnelle
  bilinéaire 114
  convexe 113
  quadratique 174
  variationnelle 58
Formule des petites perturbations 314

Identité intégrale
  fondamentale 52
  sous forme variationnelle 82
Inégalité
  de Cauchy-Bouniakovski 55
  de Friedrichs 108
Inéquation variationnelle 111
Instabilité numérique non linéaire 358
Interpolation 126, 127, 131, 135, 146

Lemme de Kellogg 17

Matrice
  bande 146
  par blocs 92
  diagonale par blocs 173
  de Gram 63

Matrice
mal conditionnée 154
tridiagonale 92
Méthode
d'approximation séquentielle 361
de décomposition 183, 186, 191, 195
du domaine auxiliaire 105
— —, analogue matriciel 202
des élements finis 72
de factorisation 216
— incomplète de Bouléev 218
de Fourier 25
de Galerkin 64
du gradient conjugué 178
de Harlow 389
de l'identité intégrale 50
itérative 150
— cyclique 151
— de Tchebycheff 163
de Ljusternik 18
des moindres carrés 67
des moments 65
de Monte-Carlo 390
de Newton 147
du plus petit résidu 177
de prédiction-correction 235
de quasi-réversibilité 391
de réduction cyclique 212
de relaxation 121
de Richardson 163
de Ritz 62
de splitting-up 239
de stabilisation 232
de surrelaxation successive 158
des variations locales 121, 122
Méthodes directes d'algèbre linéaire 392

Nombre de conditionnement 154
Norme
énergie 16
d'un opérateur 15

Opérateur
adjoint 14
auto-adjoint 14
défini positif 14
de Laplace 26
—, valeurs et fonctions propres 26
— aux différences finies 28
— — —, valeurs et vecteurs propres 28
positif 14
de projection 28
semi-défini positif 14
Opérateur
de source 35
de transition 35
Optimisation d'une méthode itérative 152

Planification des expériences 319-320
Polynôme de Lagrange 78
Problème
conditionnellement bien posé 295
d'évolution 303
inverse 295
Prolongement régulier 133

Quasi-solution d'un système d'équations linéaires 316

Rayon spectral 16
Régularisation 391

Schéma aux différences
absolument stable 40, 44
conditionnellement stable 44
de Crank-Nicholson 227
de la croix 349
à 2 cycles 367
à deux niveaux 35
explicite 34
implicite 35
numériquement stable 40
stable 41, 45
Schéma de Landau, Meiman et Khalatnikov 358
Solution généralisée d'un système 198
Symbole de Kronecker 75

Théorème
de convergence 46
de Gershgorin 21
de réciprocité pour les fonctions de Green 313
Transformation
d'Abel 139
de Fourier rapide 208

Variation première 57
Vecteur erreur 151
Vecteur résidu 151
Viscosité artificielle 354
Vitesse de convergence asymptotique d'une méthode itérative 154

TABLE DES MATIÈRES

Préface 5

Introduction 7

Chapitre premier

THÉORIE DES SCHÉMAS AUX DIFFÉRENCES. GÉNÉRALITÉS

1.1. Opérateurs et leurs adjoints 13
1.1.1. Estimation des normes de matrices (17). 1.1.2. Recherche des bornes du spectre d'une matrice positive (18). 1.1.3. Valeurs propres et fonctions propres du laplacien (26). 1.1.4. Valeurs propres et vecteurs propres du laplacien aux différences finies (28).
1.2. Approximation 31
1.3. Stabilité numérique 38
1.4. Un théorème de convergence 46

Chapitre 2

MÉTHODES DE CONSTRUCTION DE SCHÉMAS AUX DIFFÉRENCES POUR LES ÉQUATIONS DIFFÉRENTIELLES

2.1. Méthode de l'identité intégrale 50
2.2. Méthodes variationnelles en physique mathématique 56
2.2.1. Méthode de Ritz (62). 2.2.2. Méthode de Galerkin (64). 2.2.3. Méthode des moindres carrés (67).
2.3. Schémas variationnels aux différences pour les équations à coefficients discontinus 68
2.3.1. Construction d'équations de diffusion discrétisées simples par la méthode de Ritz (69). 2.3.2. Construction de schémas aux différences simples par la méthode de Galerkin (des éléments finis) (72).
2.4. Sur la formation de sous-espaces pour les méthodes variationnelles en dimension un 74
2.4.1. Construction de schémas variationnels aux différences de haute précision (75). 2.4.2. Construction d'une base de fonctions trigonométriques et son rôle dans les méthodes variationnelles (79). 2.4.3. Identité intégrale sous forme variationnelle (82).
2.5. Schémas variationnels aux différences pour une équation elliptique en dimension deux 88
2.5.1. Méthode de Ritz (88). 2.5.2. Méthode de Galerkin (94). 2.5.3. Construction de sous-espaces (98).
2.6. Méthodes variationnelles pour des problèmes à plusieurs variables 101

2.6.1. Construction de sous-espaces (101). 2.6.2. Formation de schémas variationnels aux différences point par point (103).
2.7. Méthode du domaine auxiliaire 105
2.8. Problèmes d'extrémum avec contraintes et inéquations variationnelles 111
2.8.1. Eléments de la théorie générale (112). 2.8.2. Quelques problèmes d'extrémum (114). 2.8.3. Méthodes numériques pour les problèmes d'extrémum (120).

Chapitre 3

INTERPOLATION DES FONCTIONS DISCRÈTES

3.1. Interpolation des fonctions d'une variable 127
3.1.1. Interpolation des fonctions d'une variable par les fonctions splines du troisième degré (127). 3.1.2. Interpolation cubique par morceaux avec lissage (131). 3.1.3. Prolongements réguliers (133).
3.2. Interpolation à deux et à plusieurs variables 135
3.3. Approximation r-régulière de fonctions de plusieurs variables 137
3.4. Eléments de la théorie générale des fonctions splines 143

Chapitre 4

MÉTHODES DE RÉSOLUTION DES PROBLÈMES STATIONNAIRES

4.1. Eléments de la théorie des méthodes itératives 150
4.2. Certaines méthodes itératives et leur optimisation 152
4.2.1. Méthode itérative simple (152). 4.2.2. Convergence et optimisation des méthodes itératives stationnaires (155). 4.2.3. Surrelaxation successive (158). 4.2.4. Méthode itérative de Tchebycheff (163). 4.2.5. Comparaison de la vitesse de convergence des méthodes itératives pour les systèmes d'équations aux différences (171).
4.3. Méthodes itératives non stationnaires 174
4.3.1. Des théorèmes de convergence (174). 4.3.2. Méthode du plus petit résidu (177). 4.3.3. Méthode du gradient conjugué (178).
4.4. Méthode de décomposition 183
4.4.1. Cas commutatif (186). 4.4.2. Cas non commutatif (191). 4.4.3. Optimisation des méthodes de décomposition au sens de Tchebycheff et du Calcul des variations (195).
4.5. Méthodes itératives pour les systèmes à matrices dégénérées . . 197
4.5.1. Un système compatible (198). 4.5.2. Un système incompatible (200). 4.5.3. Analogue matriciel de la méthode du domaine auxiliaire (202).
4.6. Méthodes itératives pour les problèmes à entrées inexactes . . . 205
4.7. Méthodes directes de résolution des équations aux différences finies 207
4.7.1. Transformation de Fourier rapide (208). 4.7.2. Méthode de réduction cyclique (212). 4.7.3. Factorisation des équations aux différences (215).

Chapitre 5

MÉTHODES DE RÉSOLUTION DES PROBLÈMES NON STATIONNAIRES

5.1. Schémas aux différences d'ordre d'approximation 2 à opérateurs dépendant du temps 227
5.2. Equations d'évolution non homogènes 230
5.3. Méthodes de décomposition des problèmes non stationnaires . . . 231

5.3.1. Méthode de stabilisation (232). 5.3.2. Méthode de prédiction-correction (235). 5.3.3. Méthode de splitting-up (239). 5.3.4. Certaines remarques générales (244).
5.4. Méthodes de décomposition pour $n > 2$ 244
5.4.1. Méthode de stabilisation (245). 5.4.2. Méthode de prédiction-correction (246). 5.4.3. Méthode de splitting-up en termes de schémas élémentaires (248). 5.4.4. Décomposition des problèmes quasi linéaires (253).
5.5. Méthode de splitting-up (approche générale) 254
5.6. Méthodes de résolution des équations hyperboliques 259
5.6.1. Méthode de stabilisation (259). 5.6.2. Réduction de l'équation des vibrations à un problème d'évolution (262).

Chapitre 6

AMÉLIORATION DE LA PRÉCISION DES SOLUTIONS APPROCHÉES AU SENS DE RICHARDSON

6.1. Equation différentielle ordinaire du premier ordre 268
6.2. Equation de diffusion à une dimension 273
6.2.1. Méthode de Galerkin (274). 6.2.2. Méthode aux différences (280).
6.3. Méthode de décomposition pour un problème d'évolution 286
6.4. Extrapolation de Richardson pour des problèmes à plusieurs variables 290

Chapitre 7

PROBLÈMES INVERSES (POSITION ET MÉTHODES NUMÉRIQUES)

7.1. Préliminaires 296
7.2. Problèmes d'évolution inverses et méthode de Fourier 300
7.3. Problèmes d'évolution inverses à opérateur dépendant du temps 303
7.4. Problèmes inverses posés en termes de la théorie des perturbations 309
7.4.1. Certaines questions de la théorie linéaire des mesures (309). 7.4.2. Fonctions adjointes et notion de fonction d'influence (310). 7.4.3. Théorie des perturbations pour les fonctionnelles linéaires (313) 7.4.4. Méthodes numériques de résolution des problèmes inverses et planification des expériences (315).
7.5. Théorie des perturbations et modèles non linéaires complexes . . . 320
7.5.1. Equations et leurs adjointes (321). 7.5.2. Equations adjointes et théories des perturbations (323). 7.5.3. Théorie des perturbations et problèmes non stationnaires (324). 7.5.4. Méthode spectrale et théorie des perturbations (326).

Chapitre 8

DES PROBLÈMES DE LA PHYSIQUE MATHÉMATIQUE

8.1. Equation de Poisson 328
8.1.1. Problème de Dirichlet pour l'équation de Poisson en dimension un (328). 8.1.2. Problème de Neumann en dimension un (330). 8.1.3. Equation de Poisson à deux variables (333). 8.1.4. Problème de conditions aux limites (340).
8.2. Equation de la chaleur 343
8.2.1. Problème de la chaleur en dimension 1 (343). 8.2.2. Problème de la chaleur en dimension 2 (347).

8.3. Equation des vibrations 348
8.4. Equation du mouvement 352
8.4.1. Equation du mouvement en dimension 1 (352). 8.4.2. Equation du mouvement à coefficients variables en dimension 2 (359). 8.4.3. Equation du mouvement à plusieurs variables (363).
8.5. Equation de transport non stationnaire 369

Chapitre 9

UN APERÇU DES MÉTHODES NUMÉRIQUES

9.1. Approximation, stabilité et convergence des schémas aux différences 382
9.2. Méthodes numériques pour les problèmes de la physique mathématique 384
9.3. Problèmes conditionnellement bien posés 390
9.4. Méthodes numériques d'algèbre linéaire 391
9.5. Optimisation des méthodes numériques 395
9.6. Certaines lignes d'évolution du Calcul numérique 397

Bibliographie 399

Index 425

À NOS LECTEURS

Les Editions Mir vous seraient très reconnaissantes de bien vouloir leur communiquer votre opinion sur le contenu de ce livre, sa traduction et sa présentation, ainsi que toute autre suggestion.

Notre adresse : Editions Mir, 2, Pervi Rijski péréoulok, Moscou, I-110, GSP, U.R.S.S.

Imprimé en Union Soviétique